Data format popularity in US Cloud in November 2010

Sergey Panitkin

BNL

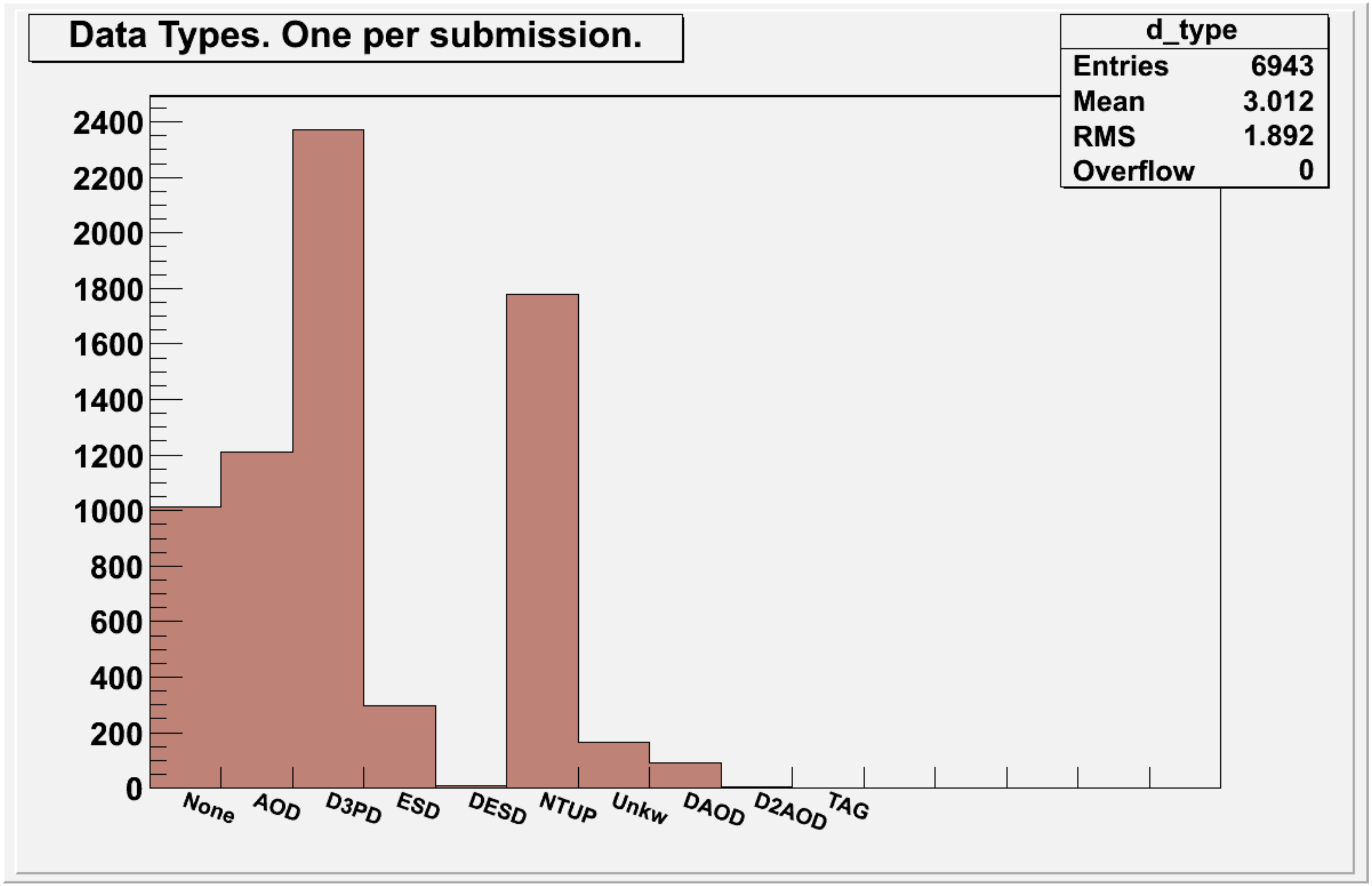
Introduction

- This is part of a study of ATLAS Grid based user analyses
- In this presentation we try to address the following questions:
    - Which data formats are typically used in ATLAS data analysis?
    - Which formats are most popular?
        - How to define popularity?
        - What is more popular format – one that used by 10 people who each submitted 1 job with 100 sub-jobs or another one that was used by 1 person who submitted 20 jobs with 50 sub-jobs each?
- The following slides will show only limited subset of available data, namely information about user analysis in US Cloud in November 2010.

Details

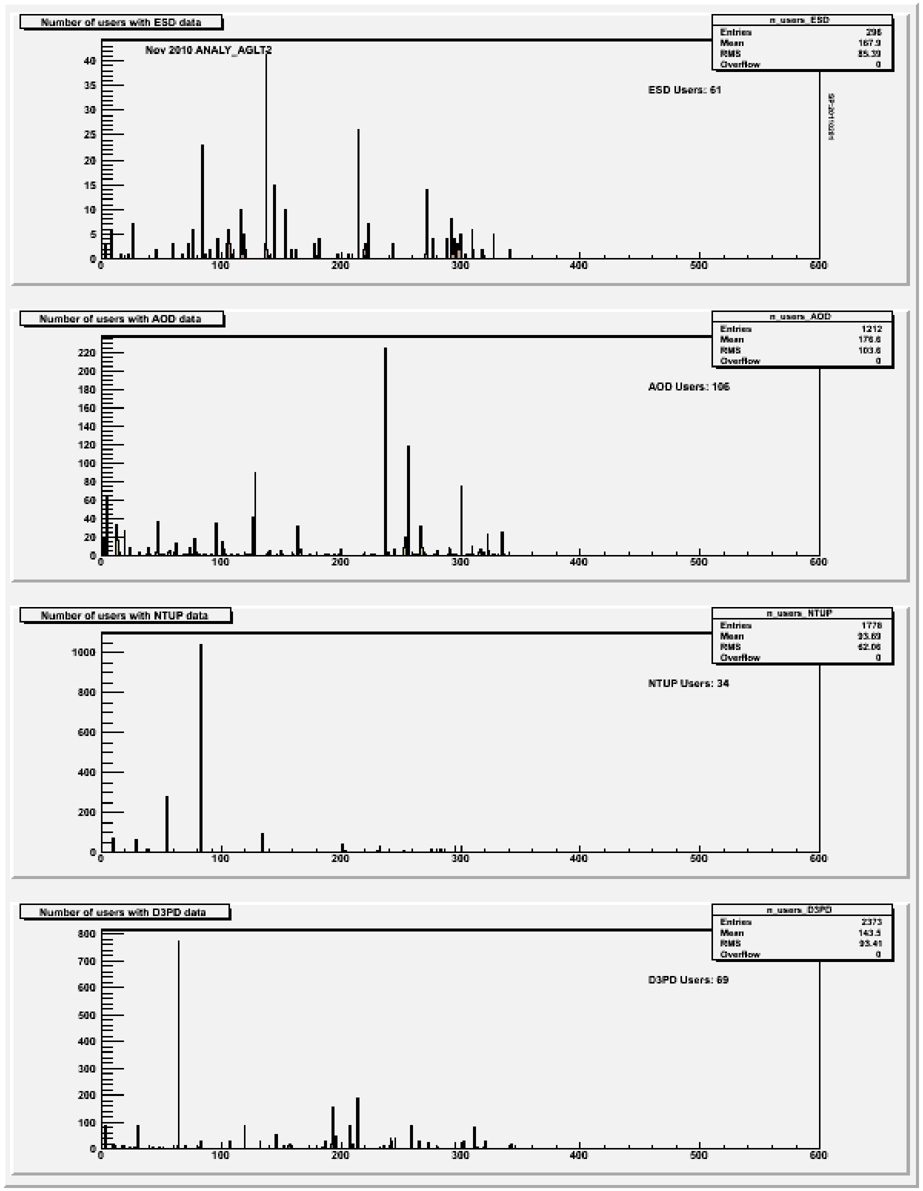
- Information was taken from Panda Oracle database
- Only user submitted (ganga, pathena, prun, pbook) analysis jobs were considered
- Only information about successfully finished analysis jobs was collected
- Input data types were identified by “prodBlock” record in Panda
- GangaRobot and HammerCloud jobs were excluded from analysis

Data types usage in AGLT2



Number of jobs (submissions) for a given input data format, **6943** jobs in total
Most jobs had D3PD and NTUP input data

Data format popularity: AGLT2



Number of jobs with a given input file format submitted per user (x-axis is arbitrary user index)

Statistics for November 2010, ANALY_AGLT2

ESD: **61** Users submitted **296** jobs with ESD input

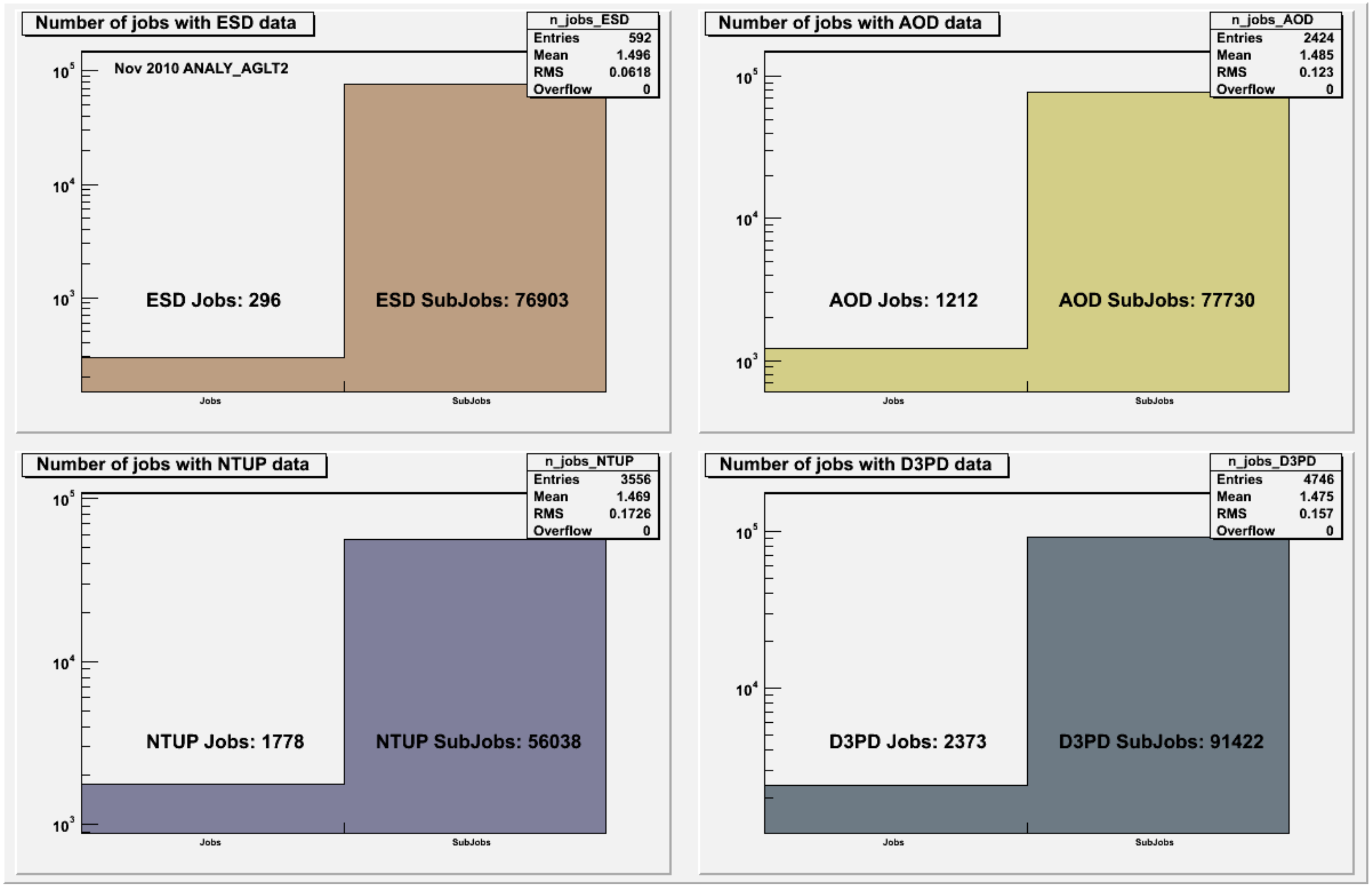
AOD: **106** Users submitted **1212** jobs with AOD input

NTUP: **34** Users submitted **1778** jobs with NTUP input

D3PD: **69** Users submitted **2373** jobs with D3PD input

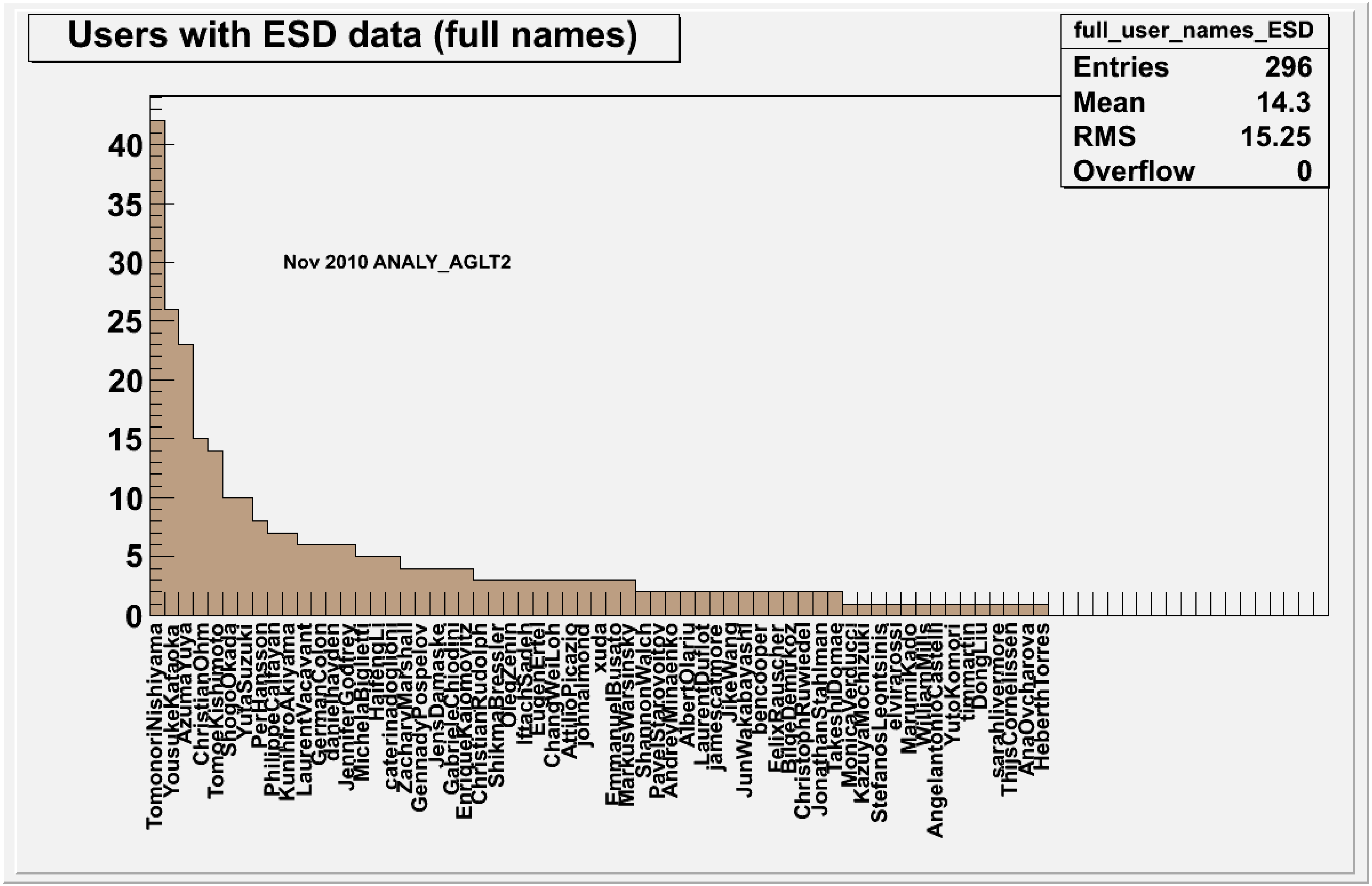
Jobs and Sub-jobs. AGLT2

Statistics for November 2010, ANALY_AGLT2



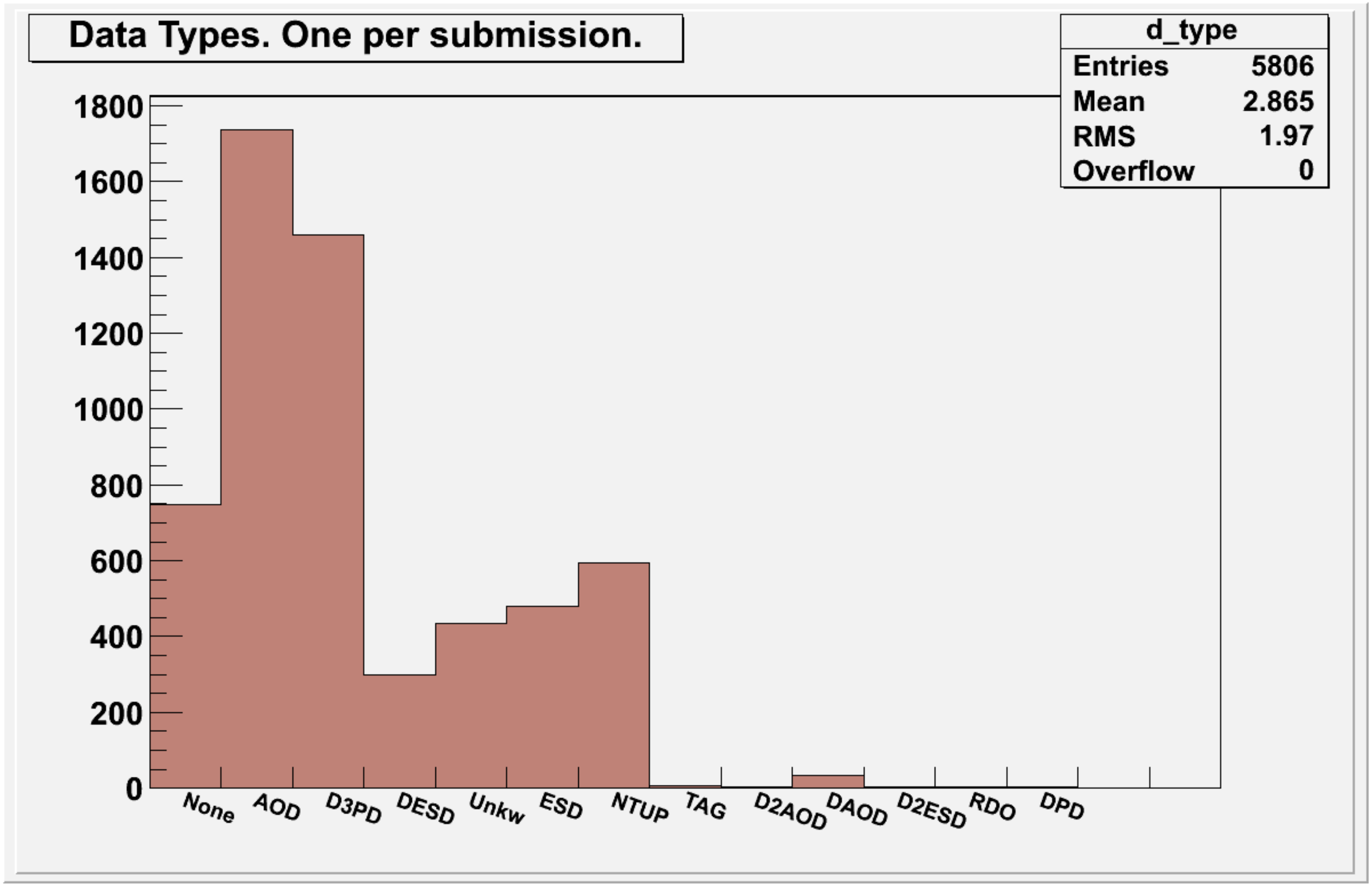
Each submitted job can have multiple sub-jobs
At AGLT2 Most jobs were submitted with D3PD and NTUP formats

ESD Users at AGLT2



Number of jobs submitted by a given user

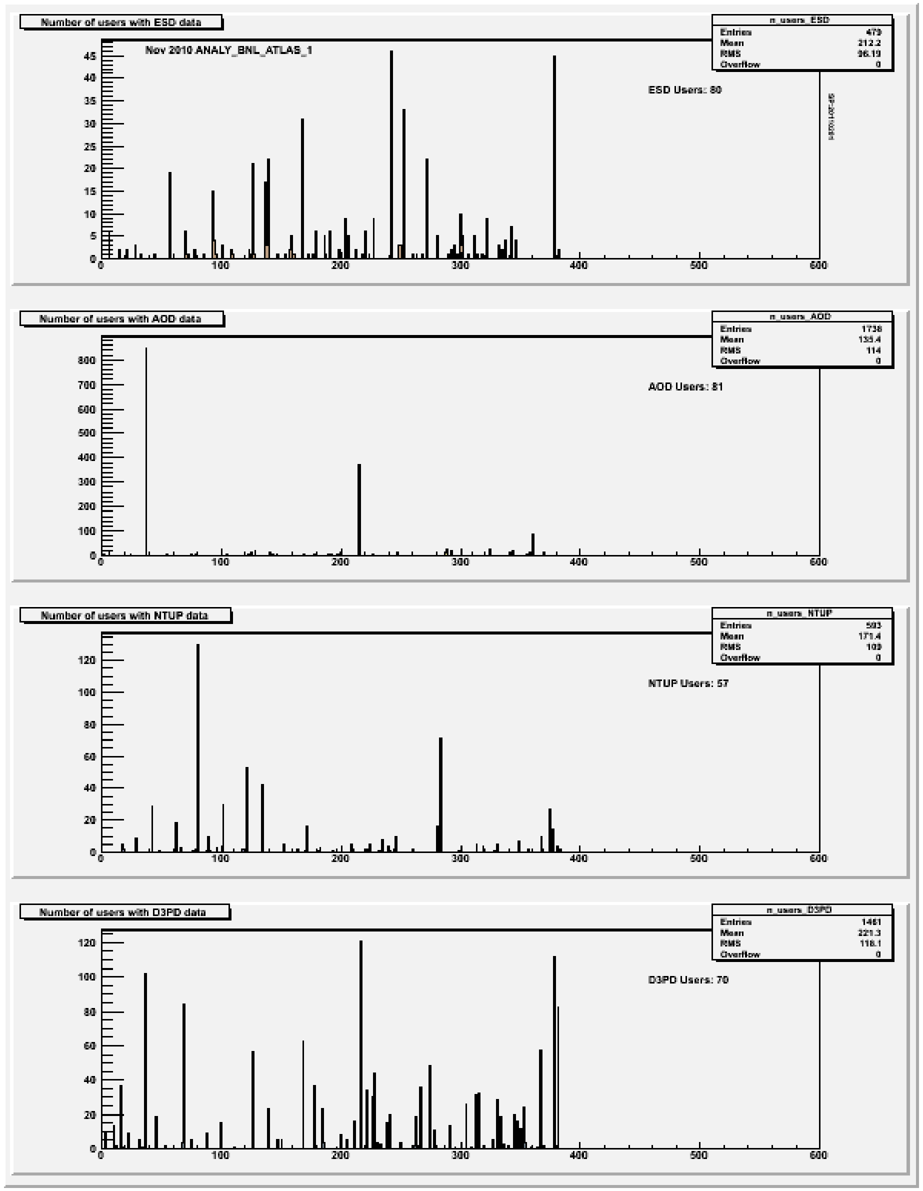
Data types usage in BNL_ATLAS_1



Number of jobs (submissions) for a given input data format, **5806** jobs in total
Most jobs had AOD and D3PD input data

Sergey Panitkin

Data format popularity: BNL_ATLAS_1



Statistics for November 2010, ANALY_BNL_ATLAS_1

ESD: **80** Users submitted **479** jobs with ESD input

AOD: **81** Users submitted **1738** jobs with AOD input

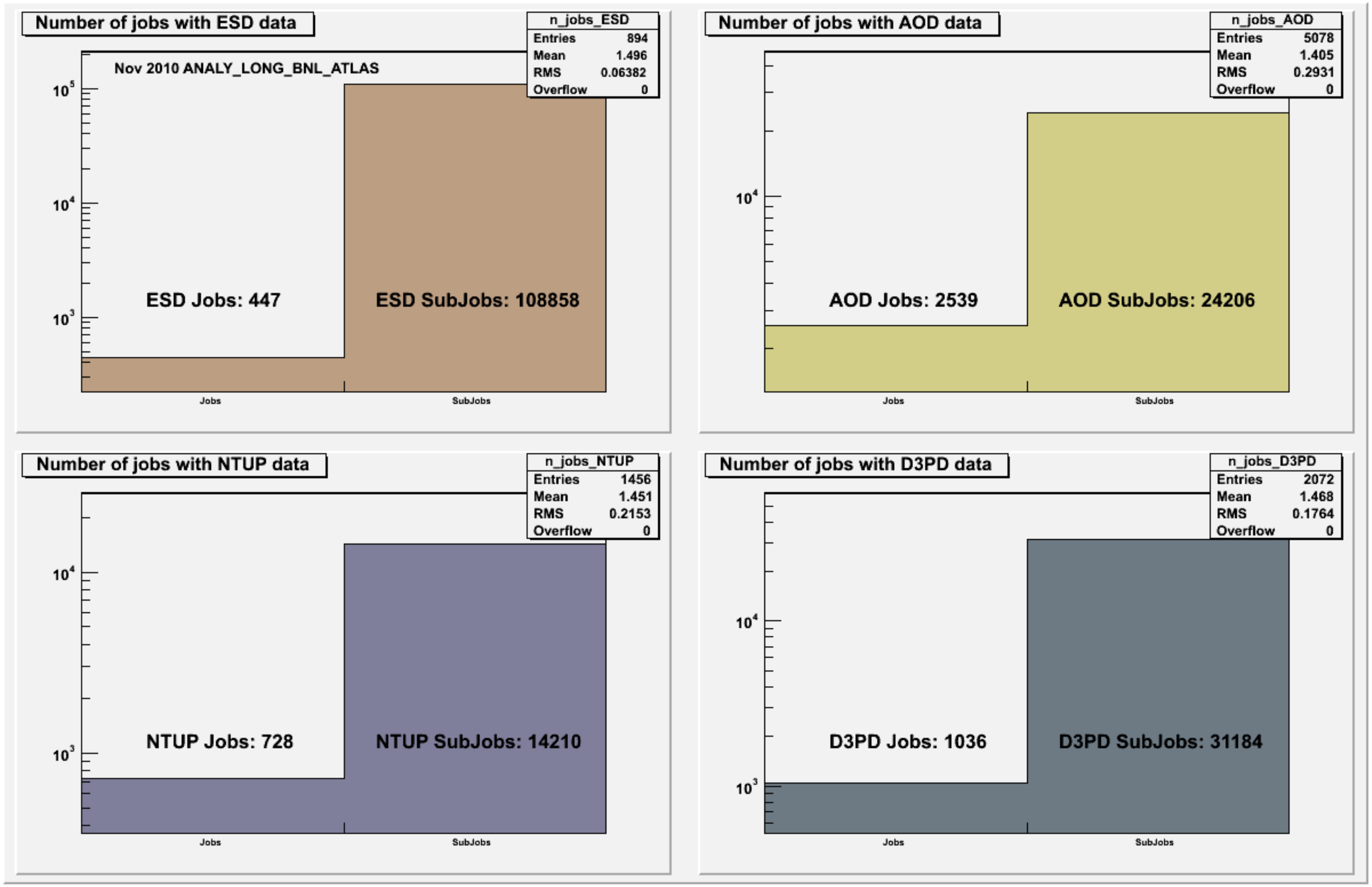
NTUP: **57** Users submitted **583** jobs with NTUP input

D3PD: **70** Users submitted **1481** jobs with D3PD input

Number of jobs with a given input file format submitted per user (x-axis is arbitrary user index)

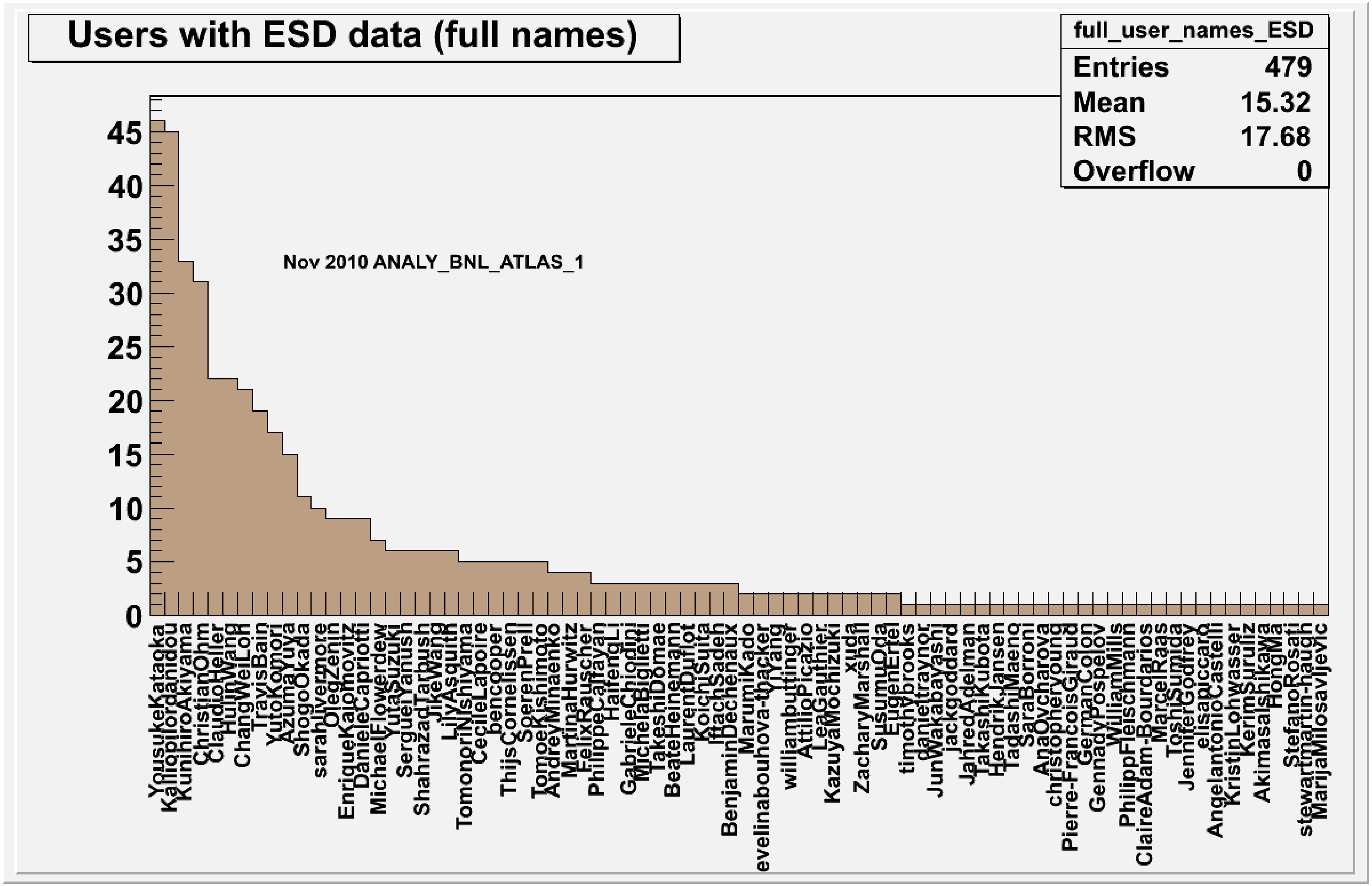
Jobs and Sub-jobs. BNL_ATLAS_1

Statistics for November 2010, ANALY_BNL_ATLAS_1



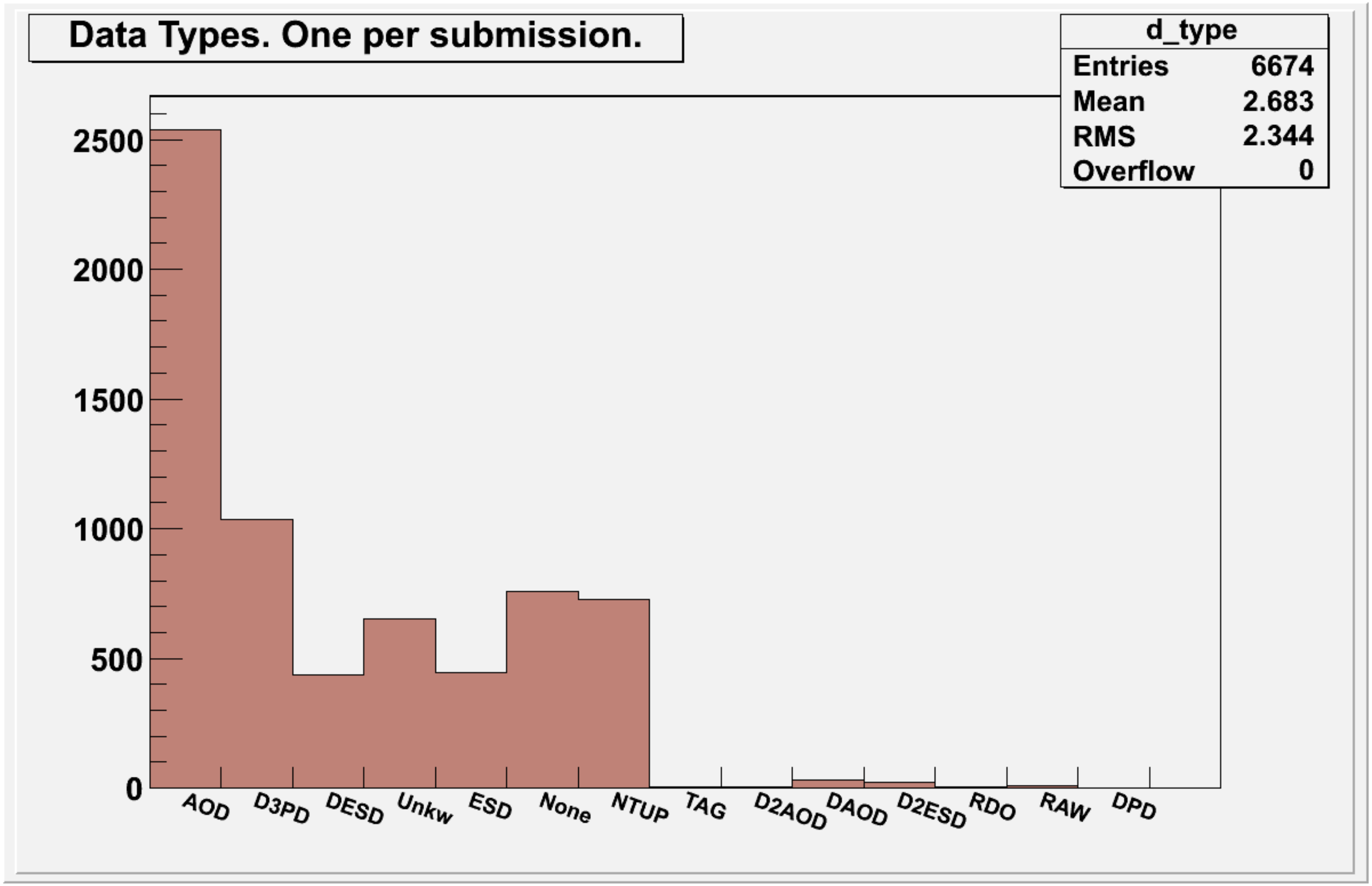
Each submitted job can have multiple sub-jobs
Jobs with ESD input had most sub-jobs , followed by D3PD, NTUP , AOD

ESD Users at BNL_ATLAS_1



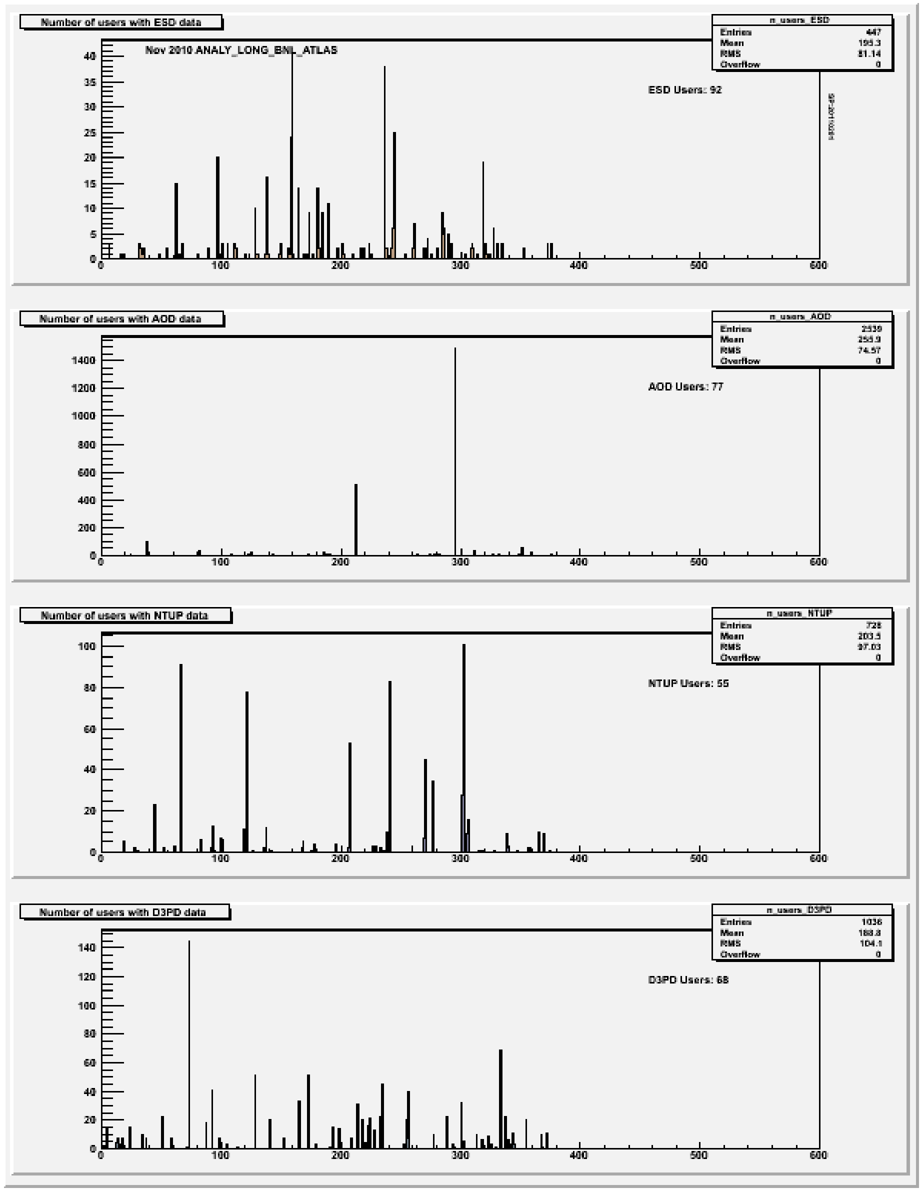
Number of jobs submitted by a given user

Data types usage in ANALY_LONG_BNL_ATLAS



Number of jobs (submissions) for a given input data format, **6674** jobs in total
Most jobs had AOD and D3PD input data

Data format popularity: LONG_BNL_ATLAS



Statistics for November 2010, ANALY_LONG_BNL_ATLAS

ESD: **92** Users submitted **447** jobs with ESD input

AOD: **77** Users submitted **2539** jobs with AOD input
AOD submission was dominated by 3 users

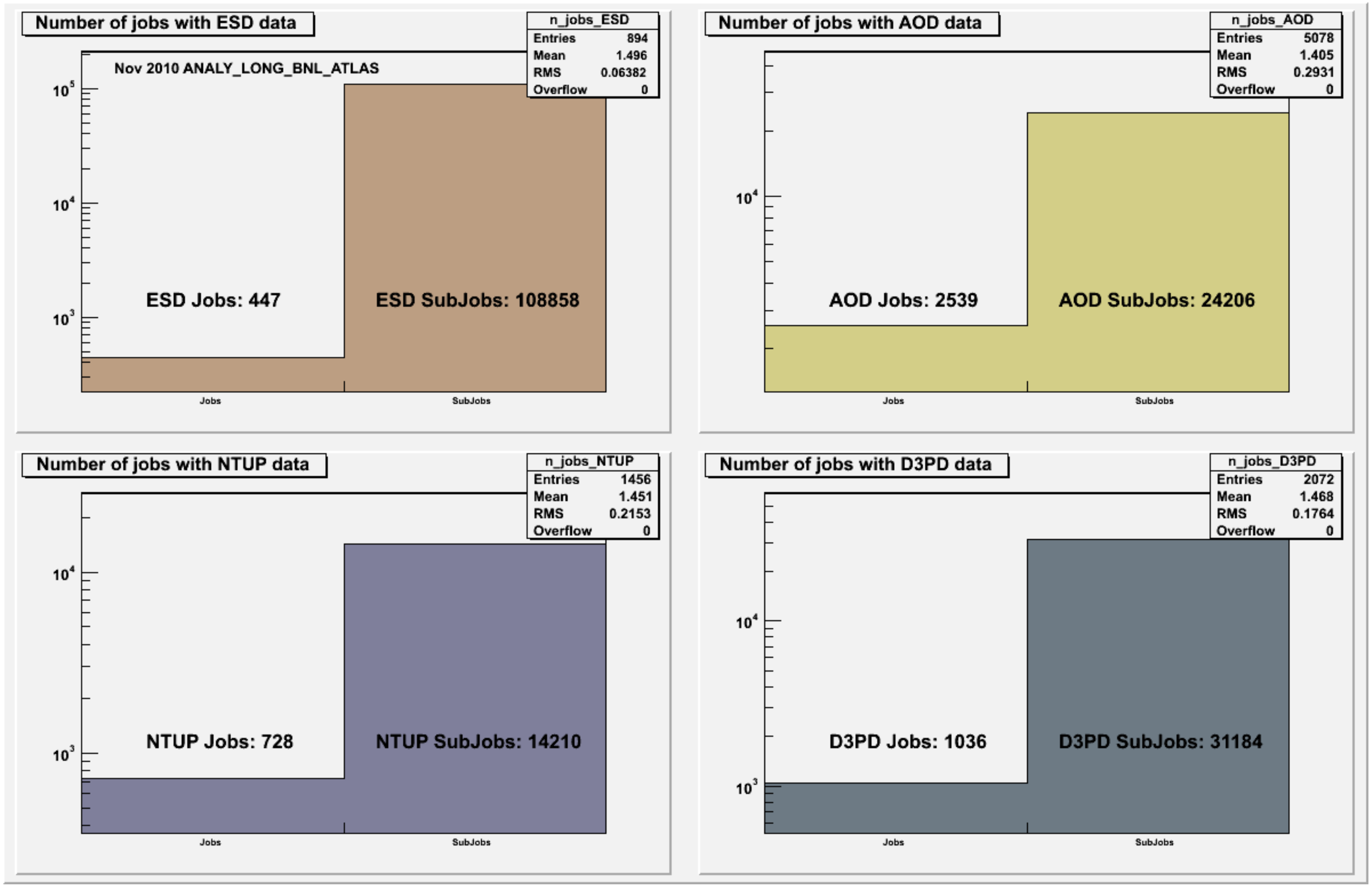
NTUP: **55** Users submitted **728** jobs with NTUP input

D3PD: **68** Users submitted **1036** jobs with D3PD input

Number of jobs with a given input file format submitted per user (x-axis is arbitrary user index)

Jobs and Sub-jobs. LONG_BNL_ATLAS

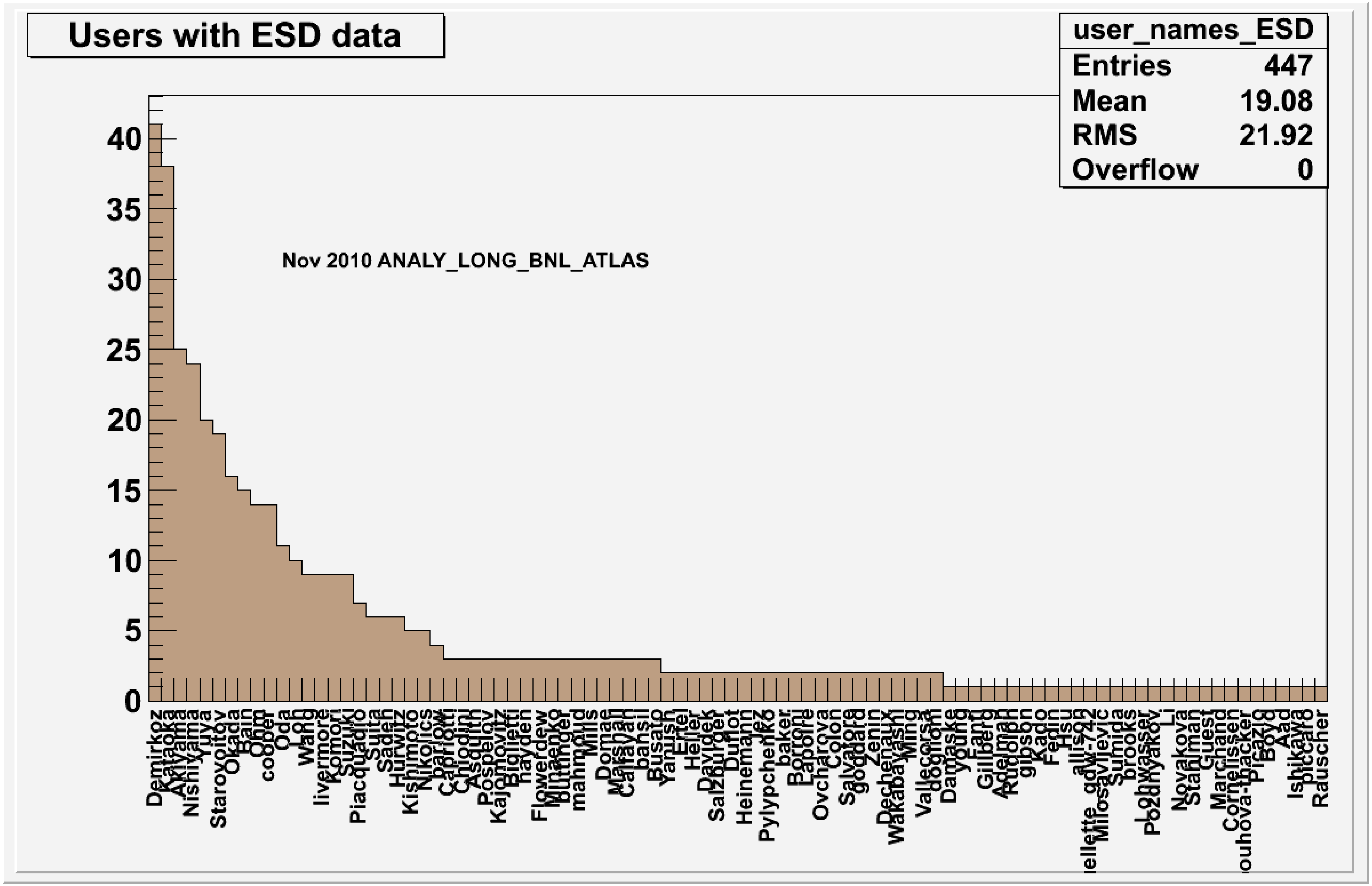
Statistics for November 2010, ANALY_LONG_BNL_ATLAS



Each submitted job can have multiple sub-jobs

Jobs with ESD input had most sub-jobs then D3PD, NTUP, AOD

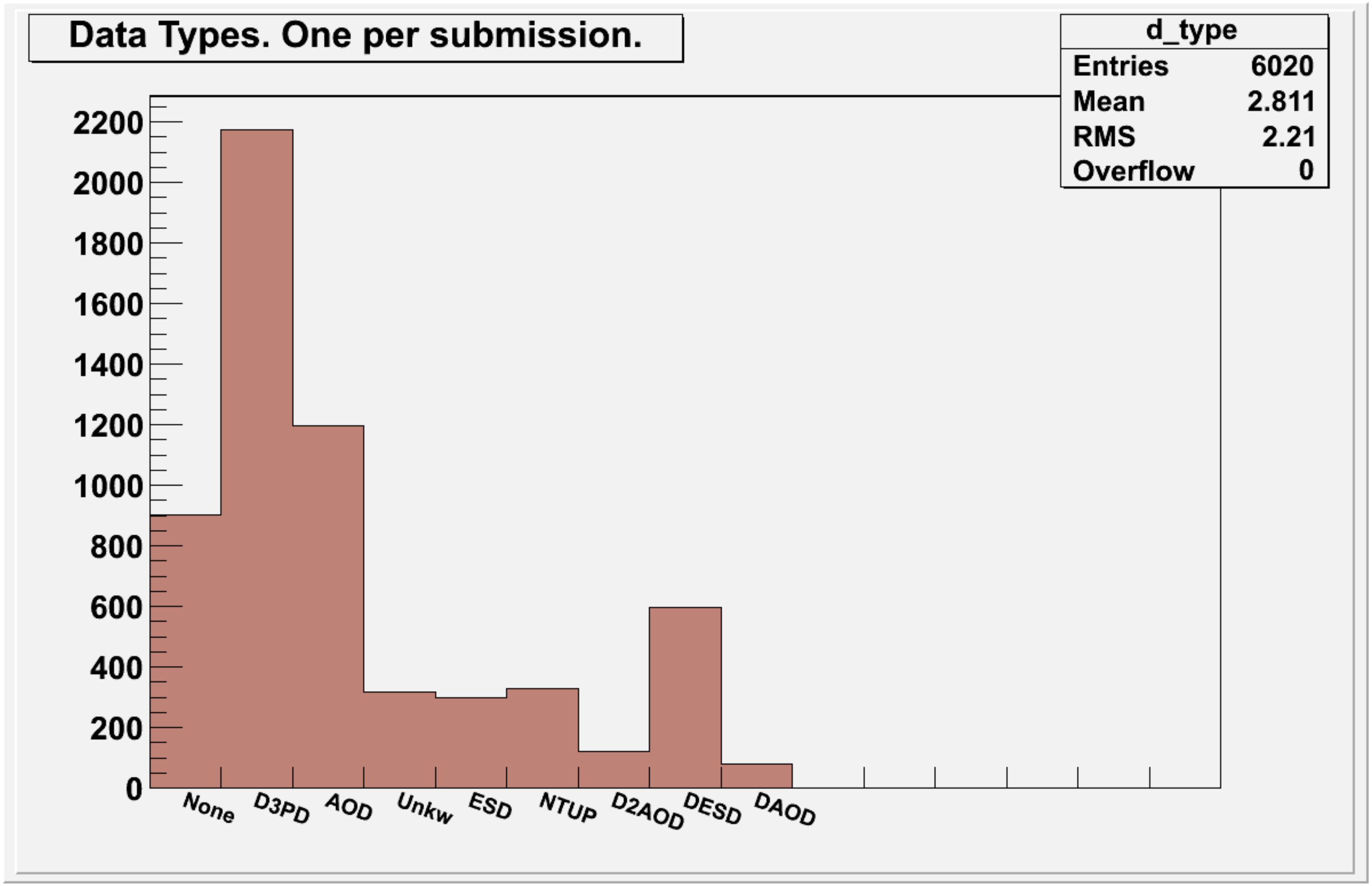
ESD Users at ANALY_LONG_BNL_1



Number of jobs submitted by a given user

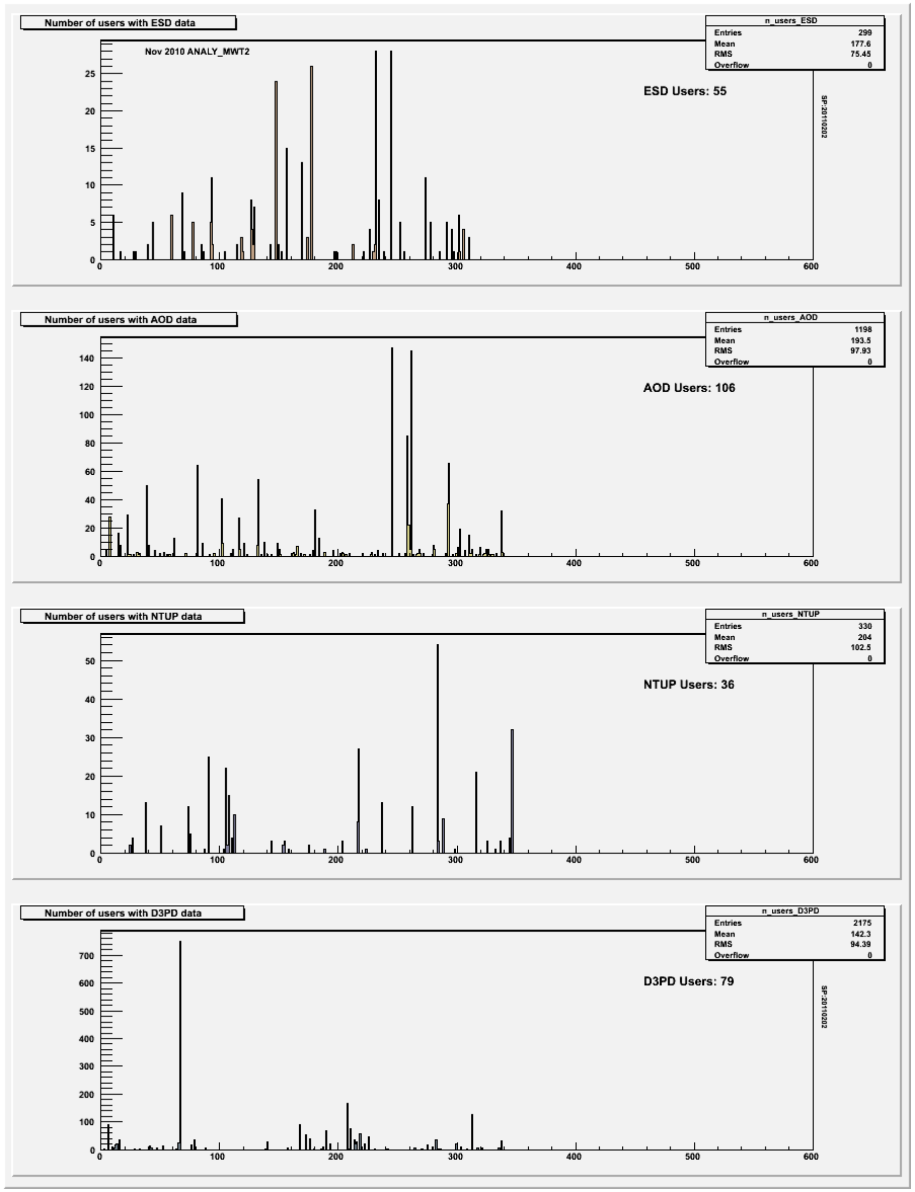
Sergey Panitkin

Data types usage in MWT2



Number of jobs (submissions) for a given input data format, **6020** jobs in total
Most jobs had AOD and D3PD input data

Data format popularity: MWT2



Statistics for November 2010, ANALY_MWT2

ESD: **55** Users submitted **299** jobs with ESD input

AOD: **106** Users submitted **1198** jobs with AOD input

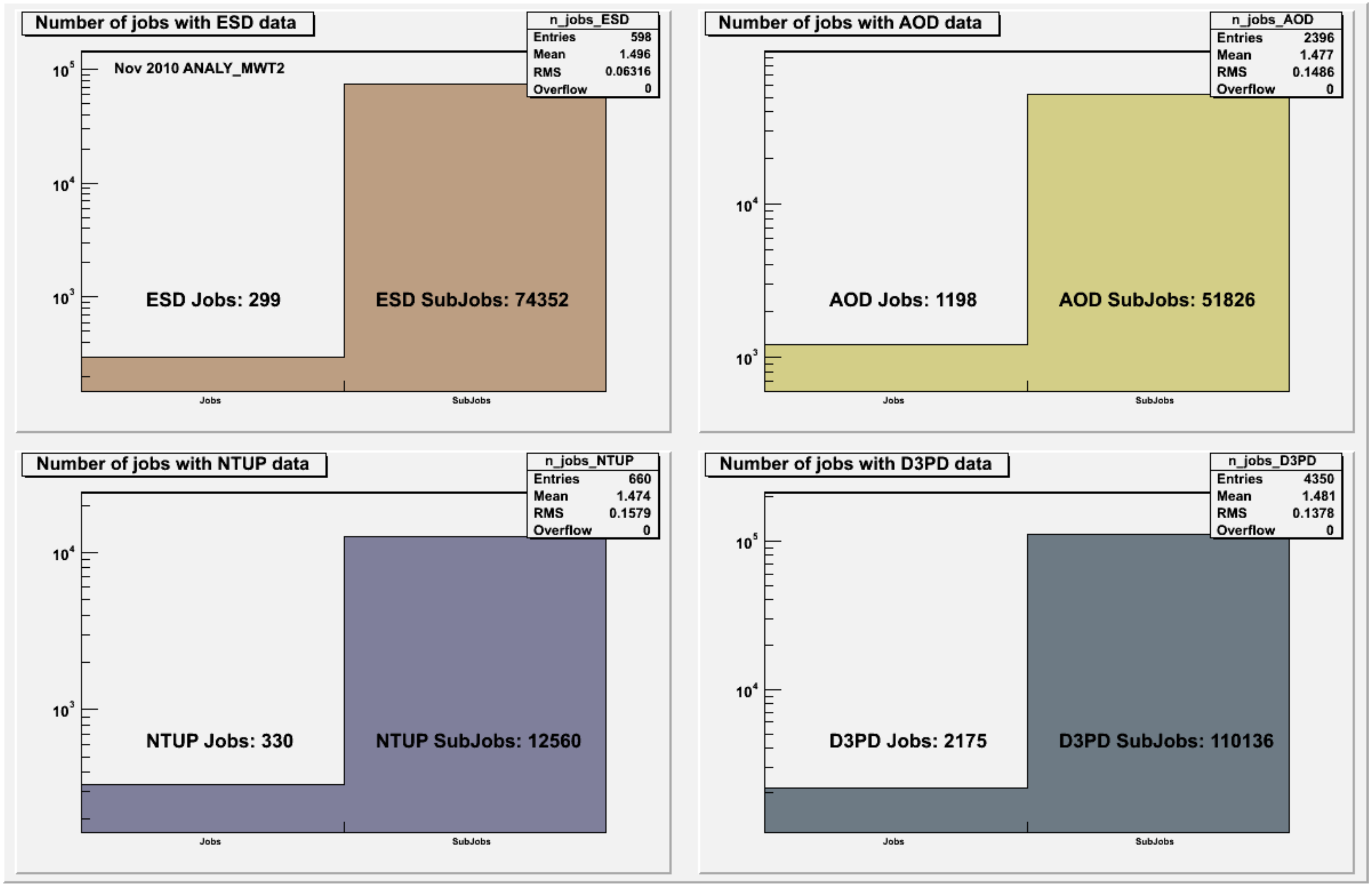
NTUP: **36** Users submitted **330** jobs with NTUP input

D3PD: **79** Users submitted **2175** jobs with D3PD input

Number of jobs with a given input file format submitted per user (x-axis is arbitrary user index)

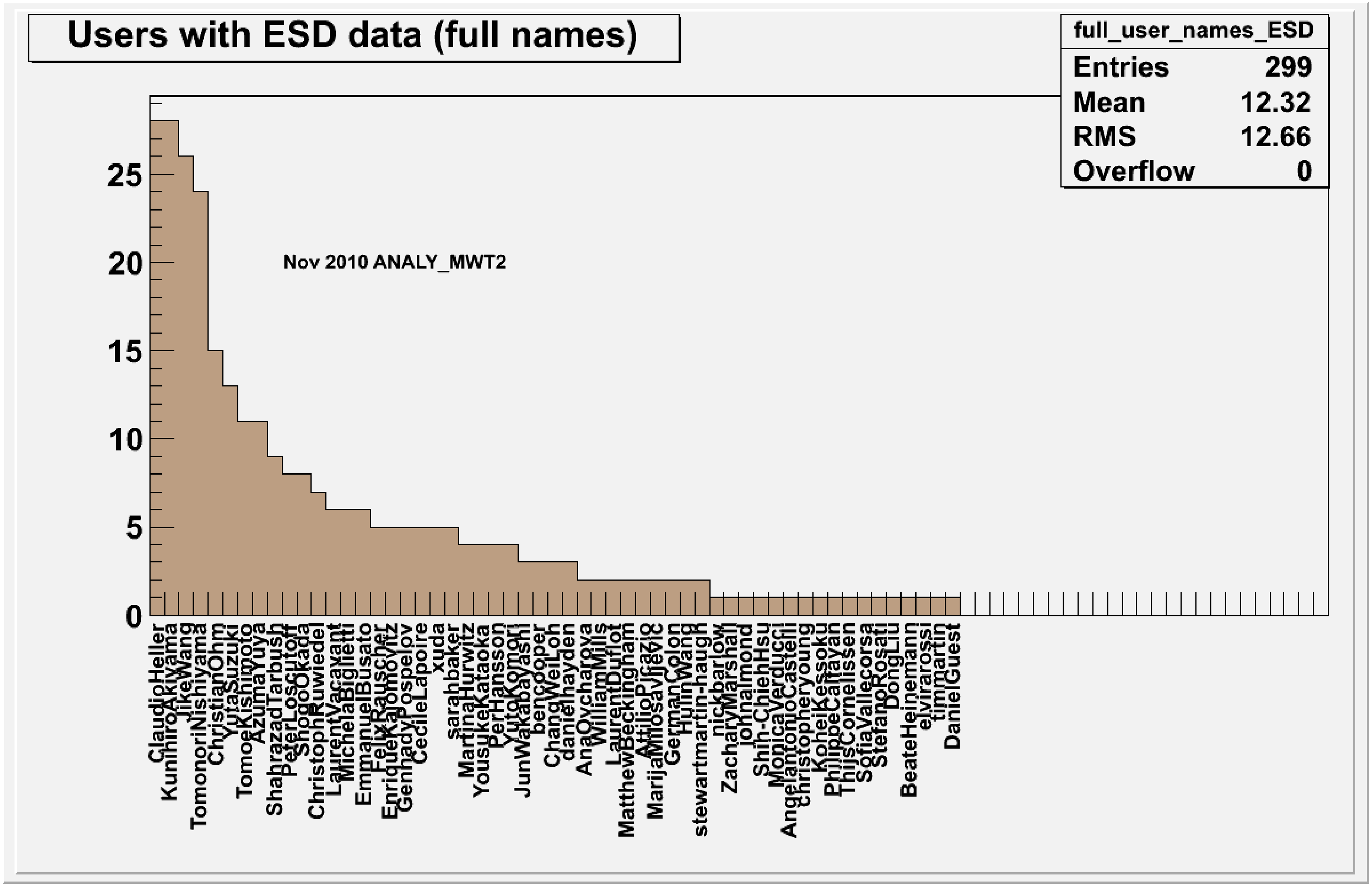
Jobs and Sub-jobs. MWT2

Statistics for November 2010, ANALY_MWT2



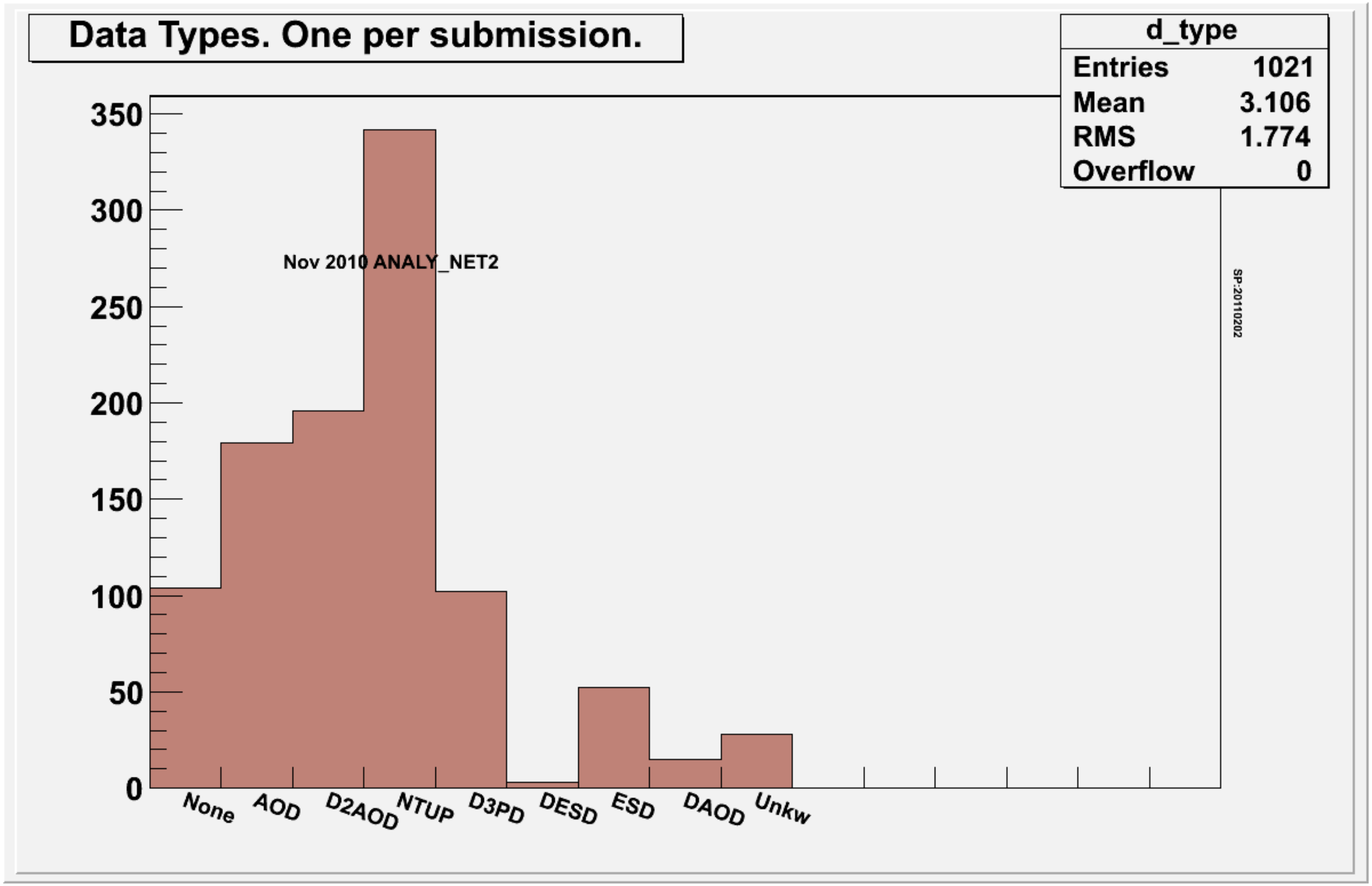
Each submitted job can have multiple sub-jobs
Jobs with ESD input had most sub-jobs ESD

ESD Users at MWT2



Number of jobs submitted by a given user

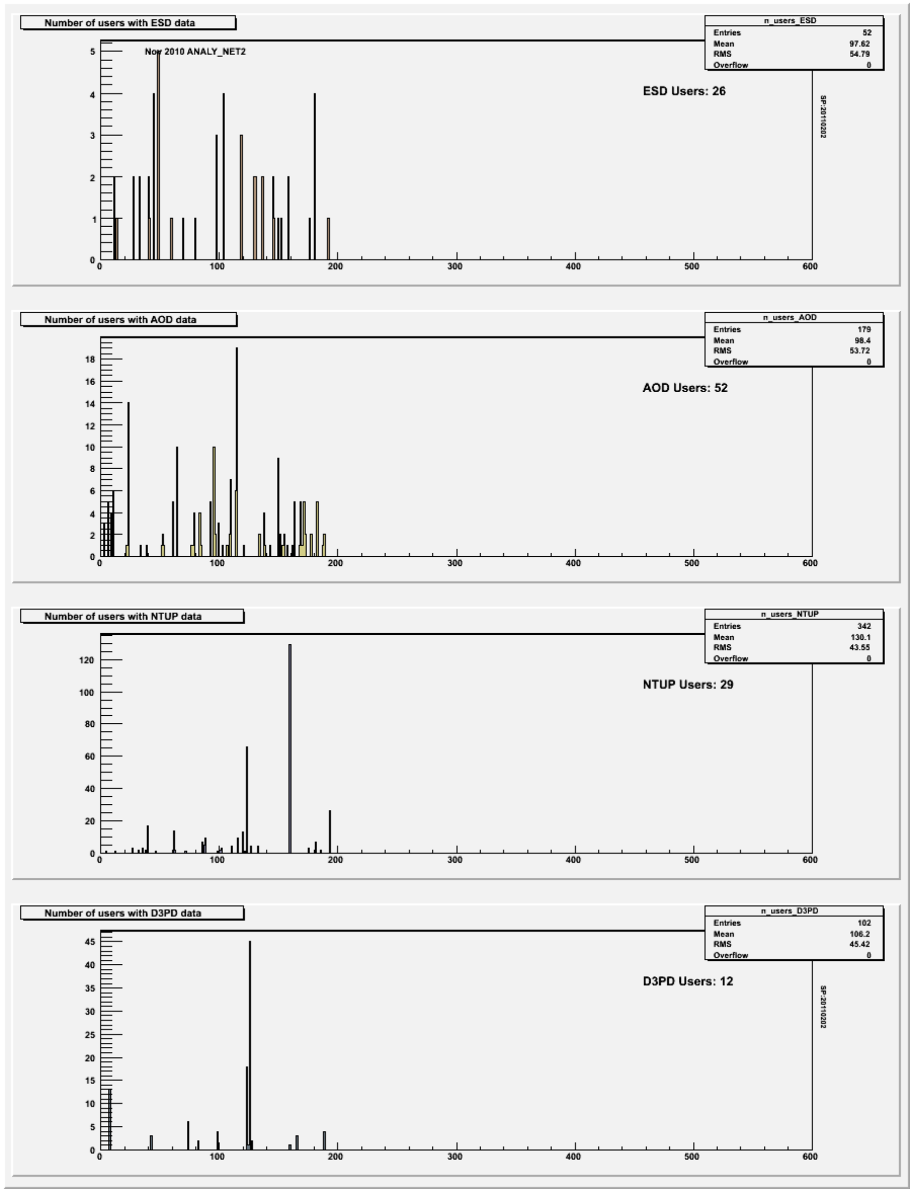
Data types usage in NET2



Number of jobs (submissions) for a given input data format, **1021** jobs in total
Most jobs had D2AOD and NTUP input data

Sergey Panitkin

Data format popularity: NET2



Statistics for November 2010, ANALY_NET2

ESD: **26** Users submitted **52** jobs with ESD input

AOD: **52** Users submitted **179** jobs with AOD input

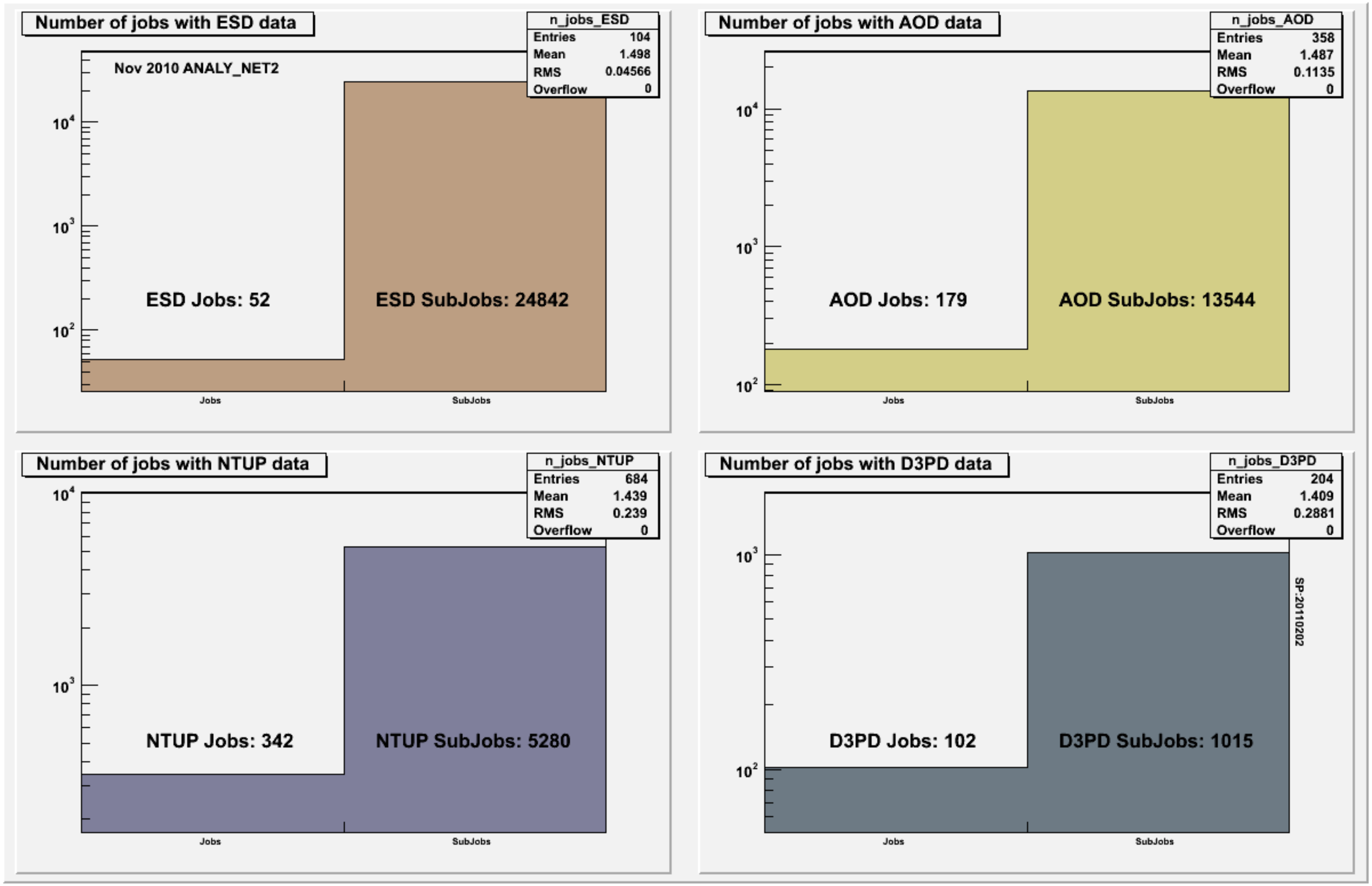
NTUP: **29** Users submitted **342** jobs with NTUP input

D3PD: **12** Users submitted **102** jobs with D3PD input

Number of jobs with a given input file format submitted per user (x-axis is arbitrary user index)

Jobs and Sub-jobs. NET2

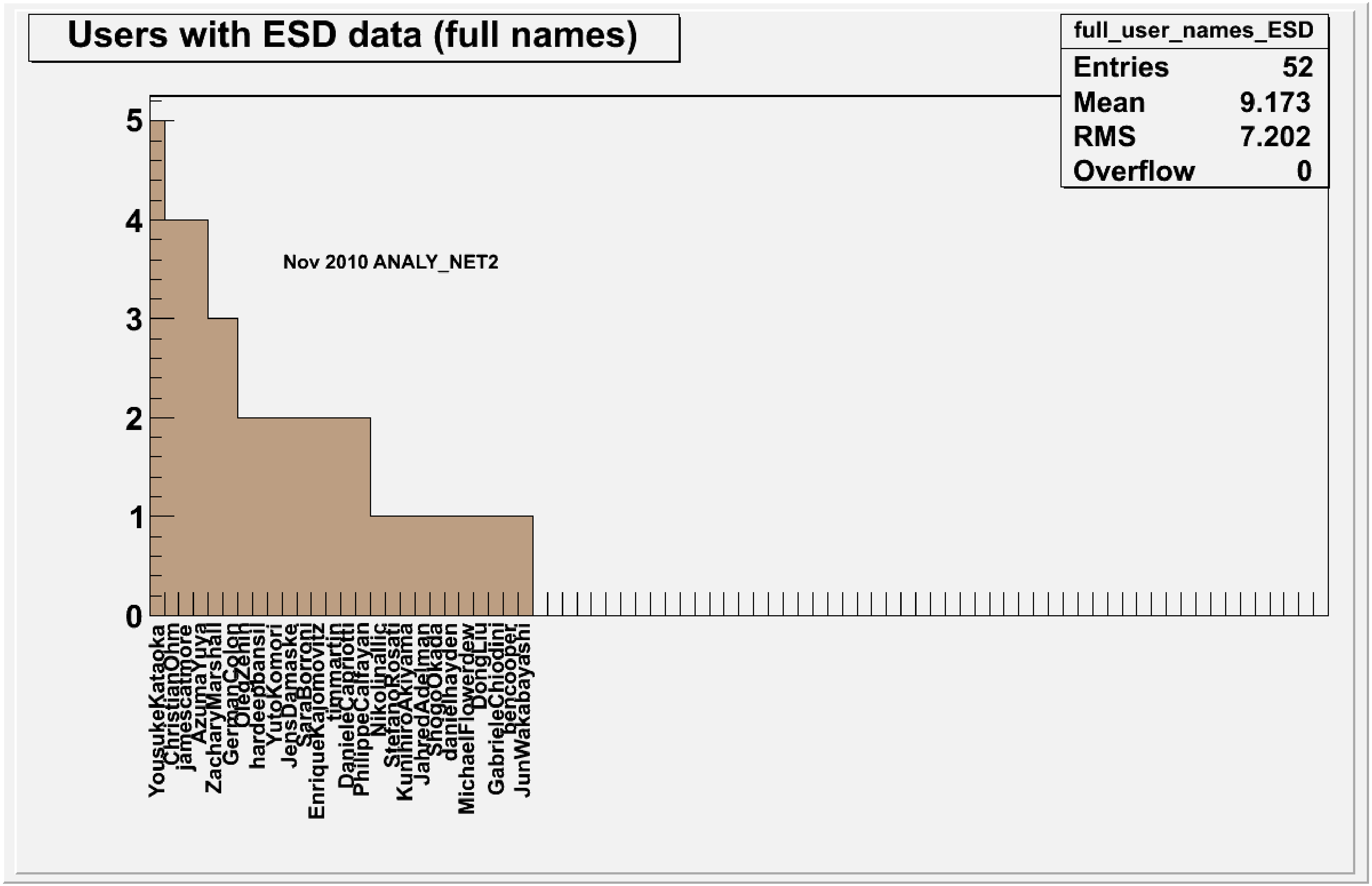
Statistics for November 2010, ANALY_NET2



Each submitted job can have multiple sub-jobs

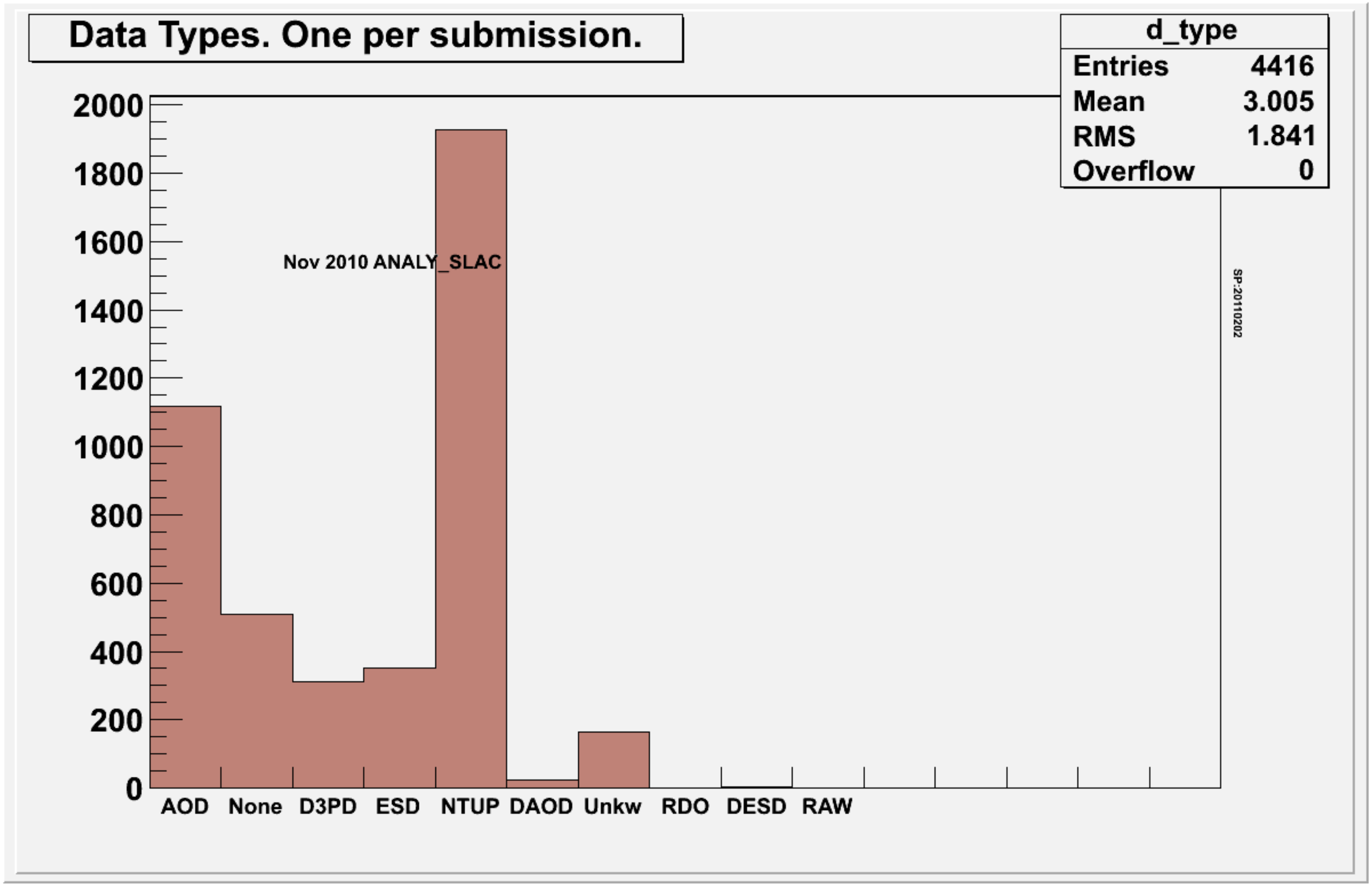
Jobs with ESD input had most sub-jobs ESD

ESD Users at NET2



Number of jobs submitted by a given user

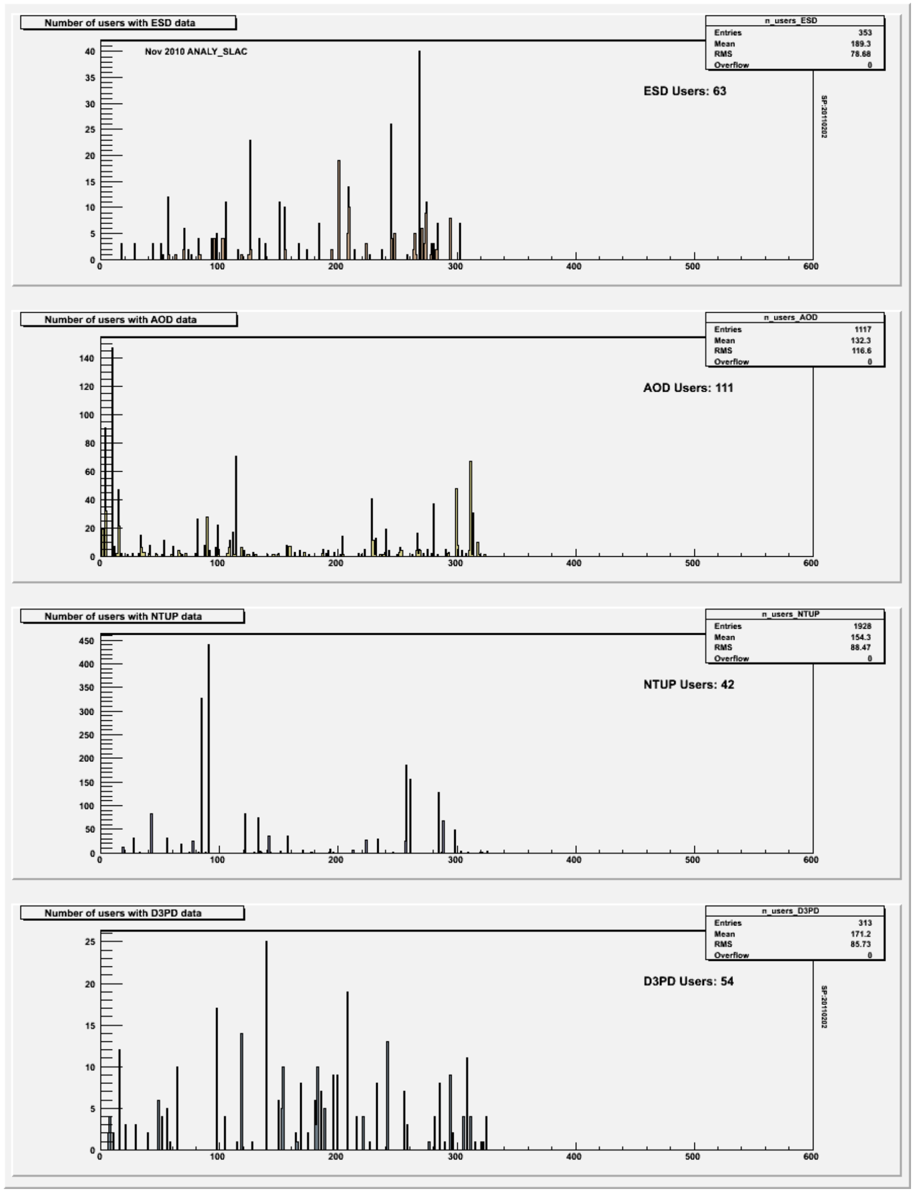
Data types usage in SLAC



Number of jobs (submissions) for a given input data format, **4416** jobs in total
Most jobs had AOD and NTUP input data

Sergey Panitkin

Data format popularity: SLAC



Statistics for November 2010, ANALY_SLAC

ESD: **63** Users submitted **353** jobs with ESD input

AOD: **111** Users submitted **1117** jobs with AOD input

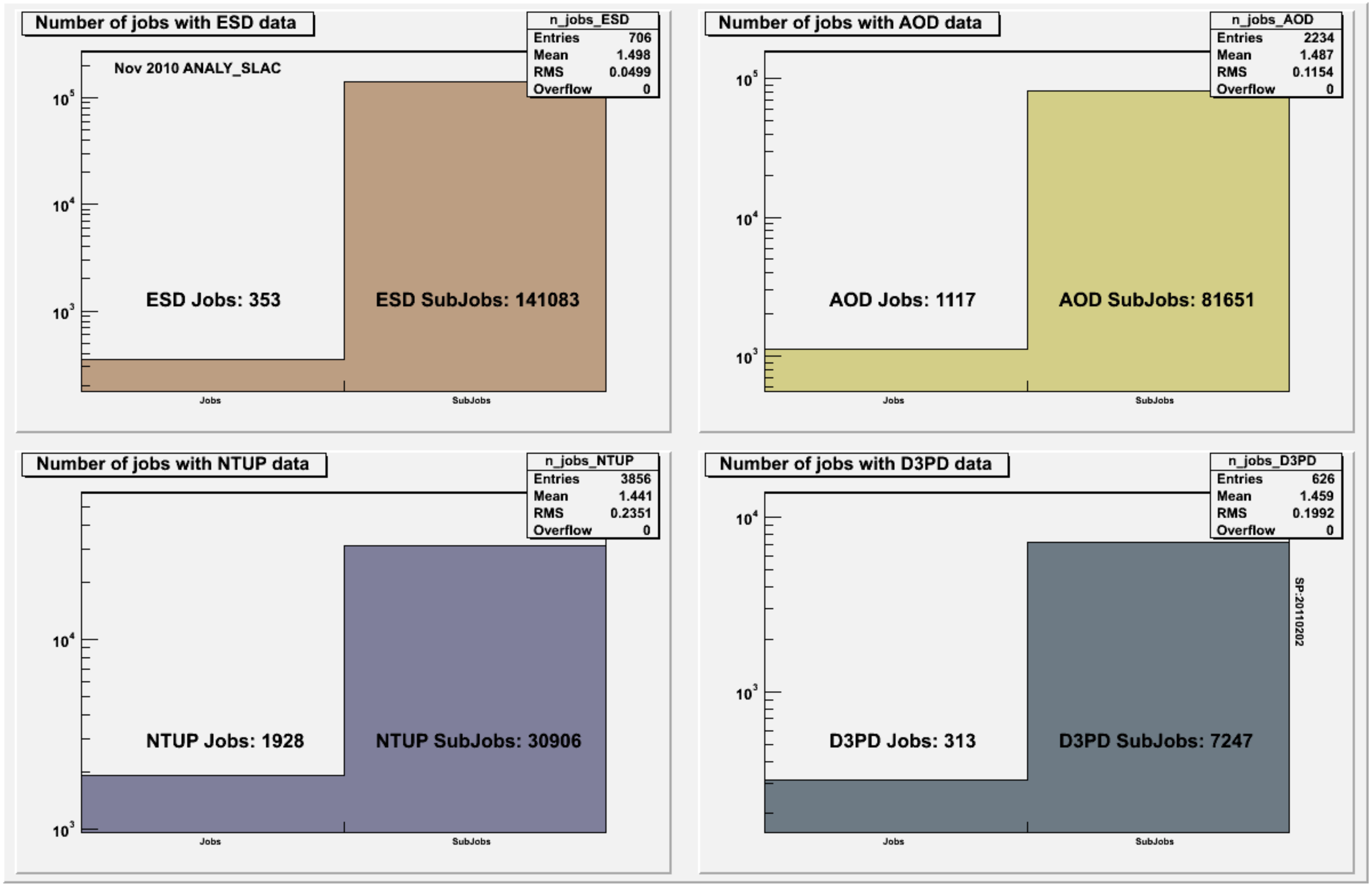
NTUP: **42** Users submitted **1928** jobs with NTUP input

D3PD: **54** Users submitted **313** jobs with D3PD input

Number of jobs with a given input file format submitted per user (x-axis is arbitrary user index)

Jobs and Sub-jobs. SLAC

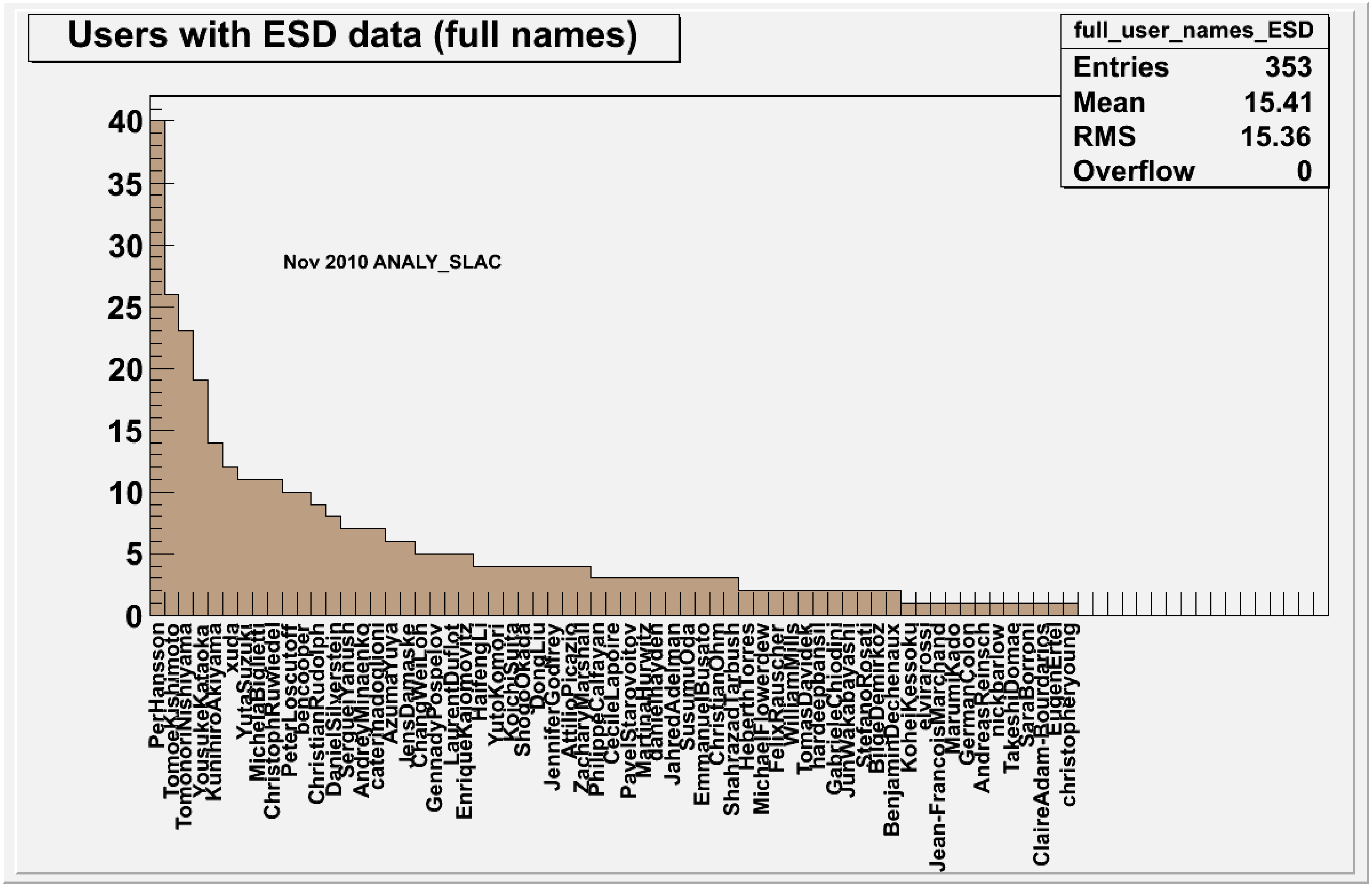
Statistics for November 2010, ANALY_SLAC



Each submitted job can have multiple sub-jobs
Jobs with ESD input had most sub-jobs

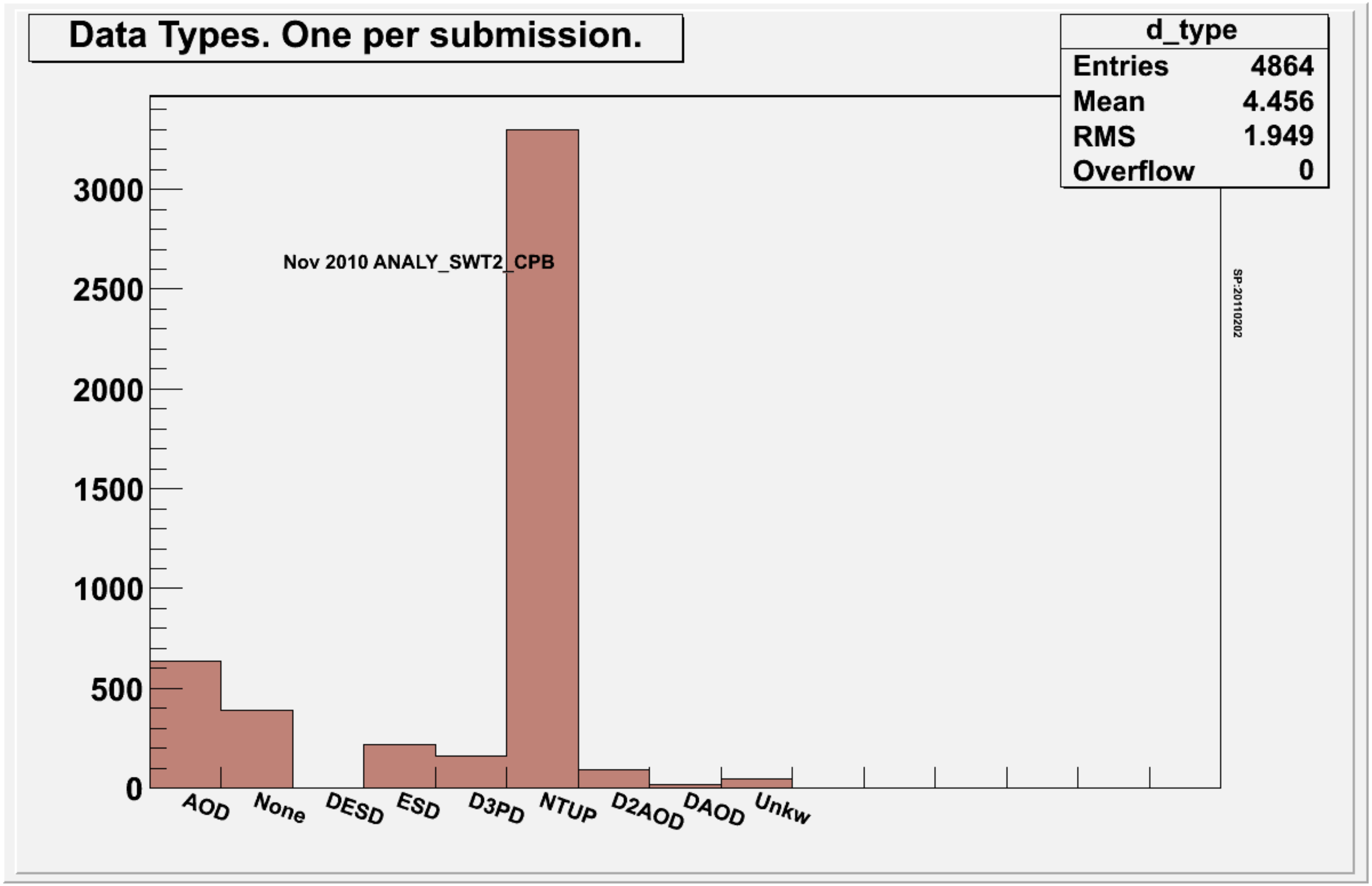
Sergey Panitkin

ESD Users at SLAC



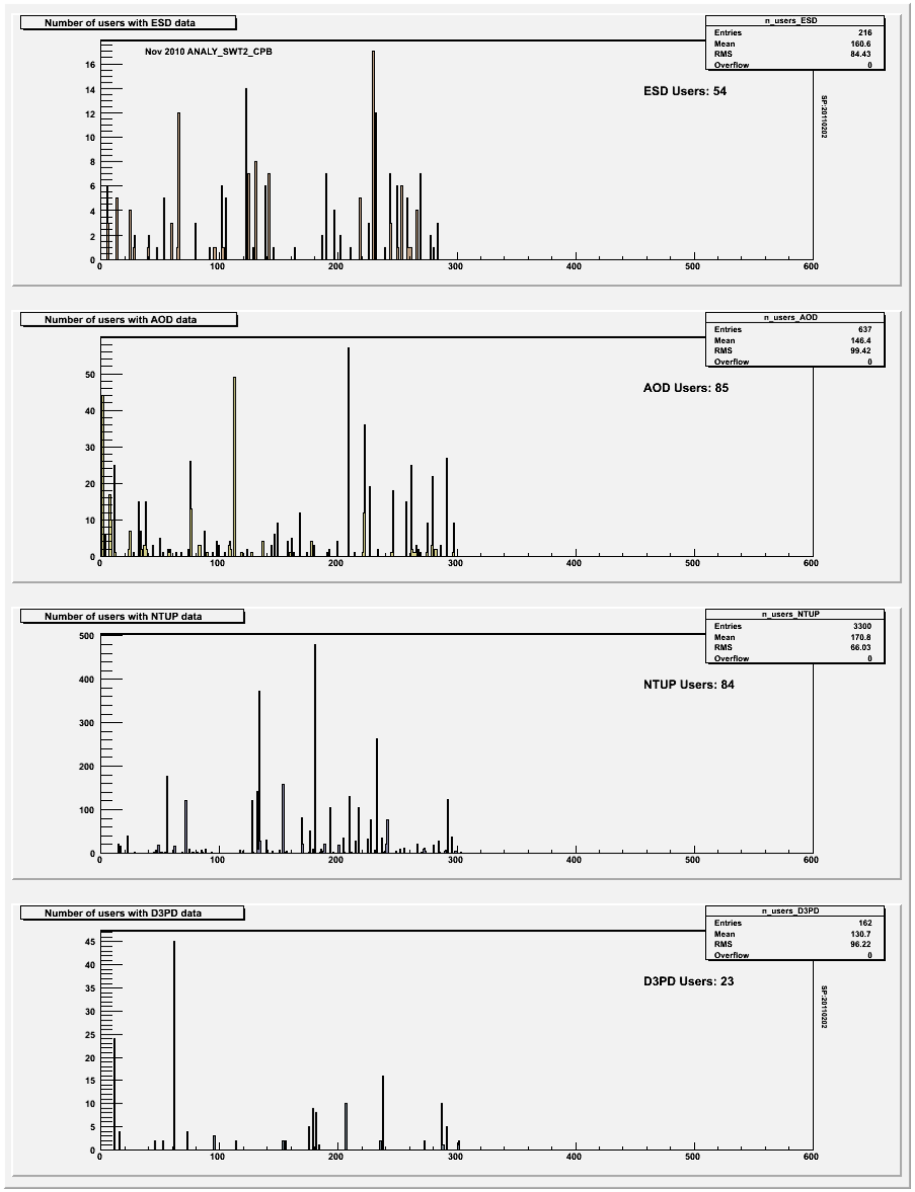
Number of jobs submitted by a given user

Data types usage in SWT2_CPB



Number of jobs (submissions) for a given input data format, **4864** jobs in total
Most jobs had NTUP input data, with AOD distant second

Data format popularity: SWT2_CPB



Statistics for November 2010, ANALY_SWT2_CPB

ESD: **54** Users submitted **216** jobs with ESD input

AOD: **85** Users submitted **637** jobs with AOD input

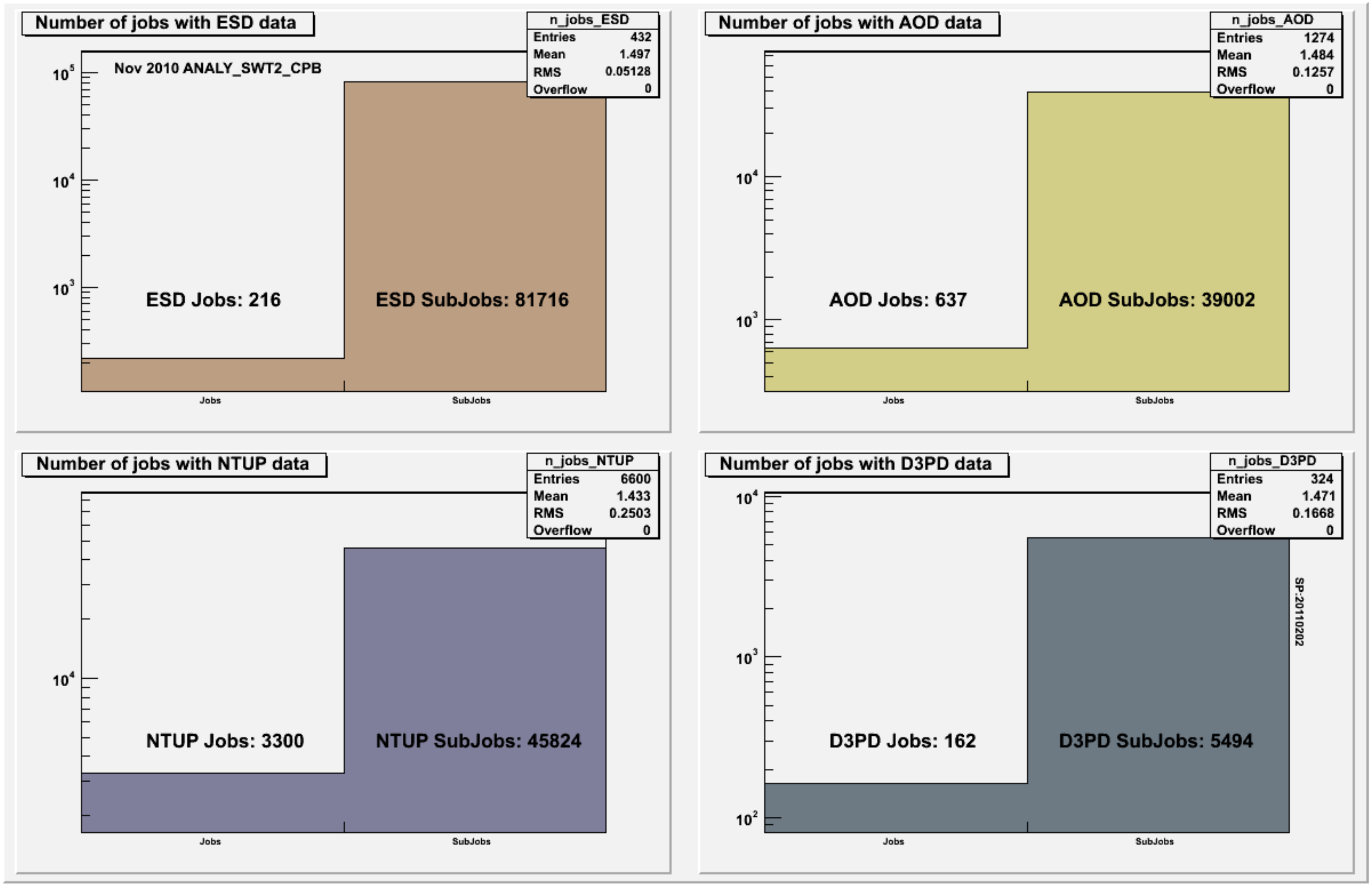
NTUP: **84** Users submitted **3300** jobs with NTUP input

D3PD: **23** Users submitted **162** jobs with D3PD input

Number of jobs with a given input file format submitted per user (x-axis is arbitrary user index)

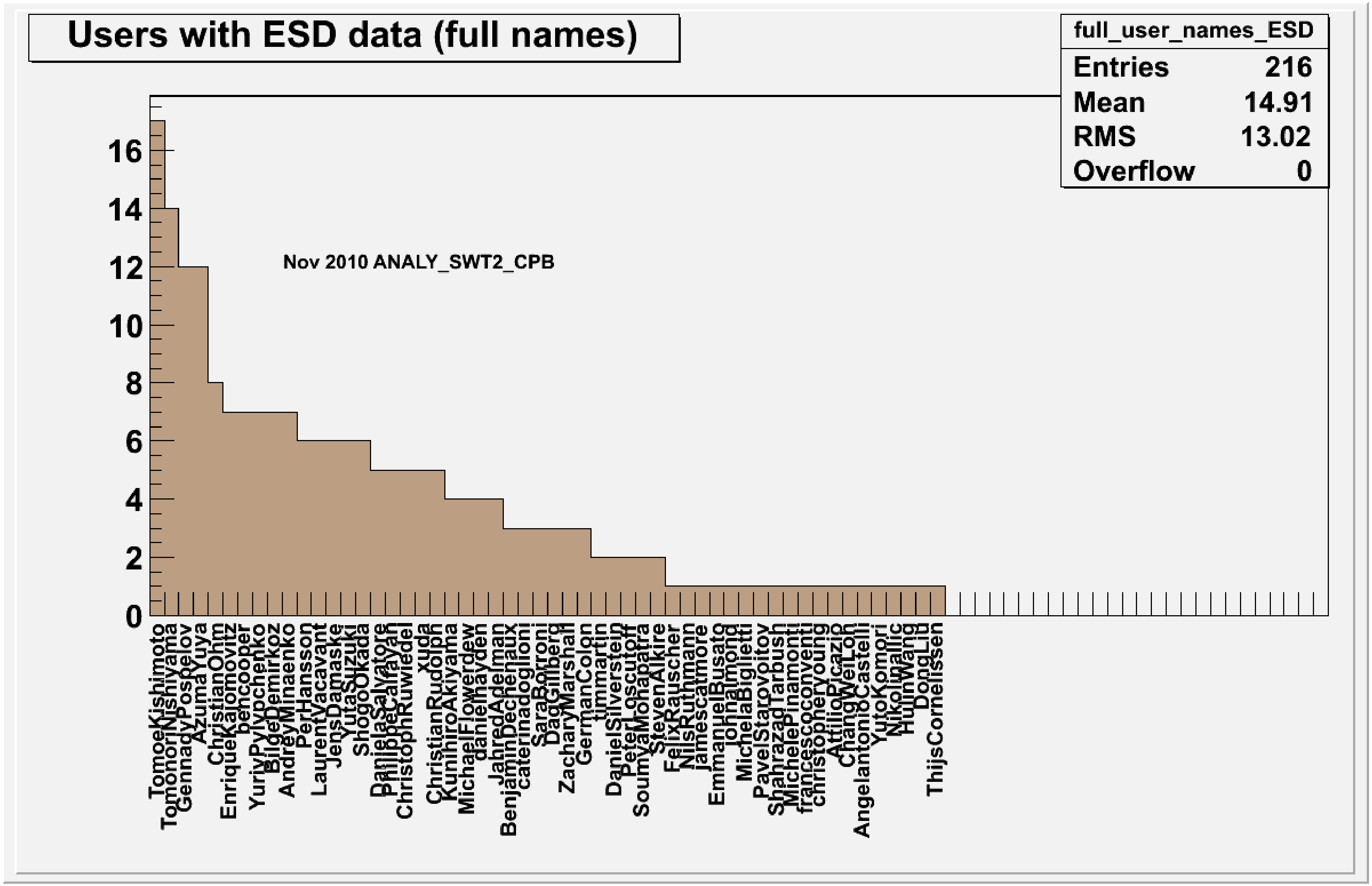
Jobs and Sub-jobs. SWT2_CPB

Statistics for November 2010, ANALY_SWT2_CPB



Each submitted job can have multiple sub-jobs
Jobs with ESD input had most sub-jobs. ESD

ESD Users at SWT2_CPB



Number of jobs submitted by a given user

Data Format Usage Summary

US Cloud, November 2010

Jobs (submissions) with a given format

	ESD	AOD	D3PD	NTUP
AGLT2	296	1212	2373	1776
BNL_1	479	1738	1461	593
BNL_LONG	447	2539	1036	728
MWT2	299	1198	2175	330
NET2	52	179	102	342
SLAC	353	1117	313	1928
SWT2	216	637	162	3300

Data format users

	ESD	AOD	D3PD	NTUP
AGLT2	61	106	69	34
BNL_1	80	81	70	57
BNL_LONG	92	77	68	55
MWT2	55	106	79	36
NET2	26	52	12	29
SLAC	63	111	54	42
SWT2	54	85	23	84

Summary

- We presented a study of usage of ATLAS data formats in user analysis on the grid
- This study was based on Panda statistics collected for November 2010, for US cloud
- For a given site and time period we can make the following observations:
    - Several hundred users were doing analysis on US cloud in November 2010.
    - About 13 input data formats were used in analyses
    - Most popular (by any definition) are ESD, AOD, D3PD and NTUP formats
    - Largest number of people analyzed AOD data.
        - This is probably the most natural definition of format popularity
    - Most analysis jobs were submitted with AOD and D3PD input
    - Small fraction (<15%) of users submits most of the jobs for any given input format
    - Most sub-jobs were using ESD and AOD input .This is most likely related to an average event size in a given format. ESD has largest event size, that forces users to split their jobs into many sub-jobs, to comply with grid site’s space and run time limits. From this point of view the sub-job based popularity metric is biased.
    - Significant number of users do analysis with derived data formats – D3PD and NTUP